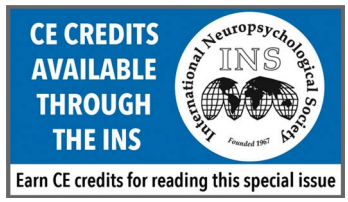
Journal of the International Neuropsychological Society (2014), **20**, 579–587.

doi:10.1017/S1355617714000307

SPECIAL SERIES – Guest Edited by Skye McDonald

Factor Structure, Convergent Validity, and Discriminant Validity of the NIH Toolbox Cognitive Health Battery (NIHTB-CHB) in Adults



Dan Mungas,[1] Robert Heaton,[2] David Tulsky,[3] Philip David Zelazo,[4] Jerry Slotkin,[5] David Blitz,[5] Jin-Shei Lai,[5] AND Richard Gershon[5]

[1]Department of Neurology, University of California, Davis, California
[2]Department of Psychiatry, University of California, San Diego, California
[3]Rusk Institute / Department of Rehabilitation Medicine, New York University, New York, New York
[4]Institute of Child Development, University of Minnesota, Minneapolis, Minnesota
[5]Department of Medical Social Sciences, Northwestern University, Chicago, Illinois

(RECEIVED September 4, 2013; FINAL REVISION March 10, 2014; ACCEPTED March 11, 2014; FIRST PUBLISHED ONLINE June 24, 2014)

Abstract

The objective of this study is to evaluate the construct validity of the NIH Neurobehavioral Toolbox Cognitive Health Battery (NIHTB-CHB) in adults. Confirmatory factor analysis was used to evaluate the dimensional structure underlying the NIHTB-CHB and Gold Standard tests chosen to serve as concurrent validity criteria for the NIHTB-CHB. These results were used to evaluate the convergent and discriminant validity of the NIHTB-CHB in adults ranging from 20 to 85 years of age. Five dimensions were found to explain the correlations among NIHTB-CHB and Gold Standard tests: Vocabulary, Reading, Episodic Memory, Working Memory and Executive Function/Processing Speed. NIHTB-CHB measures and their Gold Standard analogues defined factors in a pattern that broadly supported the convergent and discriminant validity of the NIHTB-CHB tests. This 5-factor structure was found to be invariant across 20–60 year old (N = 159) and 65–85 year old (N = 109) age groups that were included in the current validity study. Second order Crystallized Abilities (Vocabulary and Reading) and Fluid Abilities (Episodic Memory, Working Memory, Executive/Speed) factors parsimoniously explained correlations among the five first order factors. These results suggest that the NIHTB-CHB will provide both fine-grained and broad characterization of cognition across the adult age span. (*JINS*, 2014, *20*, 579–587)

Keywords: Cognition, Confirmatory factor analysis, NIH Toolbox, Adults, Construct validity, Test development

INTRODUCTION

The NIH Toolbox (NIHTB) was conceived as an initiative to develop standardized measures of cognition, emotion, motor function, and sensation that could provide common research infrastructure to facilitate integration of results across studies (Gershon et al., 2013). The cognition domain, measured by the NIH Toolbox Cognitive Health Battery (NIHTB-CHB), is the focus of this special series. The NIHTB-CHB measures multiple dimensions of cognition relevant to studies of brain and life experience determinants of cognitive function across the full range of normal cognition. The development of this battery is described in the Introduction to this special series. Briefly, a survey of knowledge leaders in adult and child cognition was used to identify cognitive sub-domains that were of highest priority. A comprehensive development process was then followed to create measures of these identified abilities that met NIHTB requirements including: (1) Applicable from age 3 to 85 years, (2) Available in English and Spanish versions), (3) Brief – the entire cognition battery can be completed in 30 min, (4) non-proprietary, and (5) based on state of the art measurement and test administration technology. This manuscript addresses the construct validity of the NIHTB-CHB.

Construct validity begins with a conceptual model that describes the expected relations between domains being

Correspondence and reprint requests to: Dan Mungas, Department of Neurology, 4860 Y Street, Suite 3700, Sacramento, CA 95817. E-mail: dmmungas@ucdavis.edu

measured and specific tests used to measure those domains. The NIHTB-CHB was designed to assess six specific domains of cognition: working memory, executive function, episodic memory, processing speed, language and reading. This test development model provides a conceptual foundation for the construct validation of the NIHTB-CHB. The measure development process also incorporated existing "gold standard" measures of these same abilities. The present "validation" study, which was conducted before the national norming study, used confirmatory factor analysis to examine the dimensional structure underlying the NIHTB-CHB and gold-standard counterparts. It also addressed the extent to which the conceptual model that guided development is reproduced in the empirical relations among NIHTB-CHB and gold standard measures.

Convergent and discriminant validity are important elements of construct validity that are based upon dimensions that explain relations among tests selected to measure different, specific domains. Construct validity is supported when (a) the empirically observed dimensions correspond to the *a-priori* conceptual model for the domains being measured, (b) individual tests are strongly related to the dimensions hypothesized from the conceptual model (convergent validity), and (c) tests are not related (or are more weakly related) to other dimensions (discriminant validity). This study examined these three elements of construct validity in relation to the NIHTB-CHB.

The broader literature on dimensions explaining correlations among tests of cognitive abilities was used to inform alternate confirmatory factor analysis models that were tested in this study. There is considerable literature on the factor structure underlying cognitive test batteries. The dimensions inherent in the Wechsler Adult Intelligence Scale and Wechsler Memory Scale have been most often studied, and factors representing verbal abilities, visual perceptual abilities, attention/concentration, and memory, sometimes involving different sub components, have frequently been found (Bowden, Carstairs, & Shores, 1999; Bowden, Cook, Bardenhagen, Shores, & Carstairs, 2004; Larrabee et al., 1983; Smith, et al., 1992; Smith, Ivnik, Malec & Tangalos, 1993). In a study of older adults using different tests, Tuokko et al. (2009) identified a three-factor model consisting of long-term retrieval, verbal abilities, and visuospatial abilities, and showed partial invariance across English and French speaking sub-groups. Mungas, Widaman, Reed, and Tomaszewski Farias (2011) identified five dimensions (episodic memory, semantic memory/language, spatial ability, attention/working memory, fluency) underlying cognitive test performance in ethnically and linguistically diverse older adults, and showed measurement invariance across Caucasian, African American, and English and Spanish speaking Hispanic subgroups. A previous publication based on the NIHTB-CHB in children showed a 5-factor solution (reading, vocabulary, episodic memory, working memory, executive/speed) in older children, ages 8–15, but a less differentiated, 3-factor solution (vocabulary, reading, fluid abilities) in younger, 3–6 year olds (Mungas, 2013).

Some studies have examined factorial structure of Executive function measures. A seminal investigation of the factor structure of EF in young adulthood used confirmatory factor analysis to extract three correlated latent variables from several commonly used EF tasks, believed to represent inhibition, working memory, and shifting (Miyake et al., 2000). Crane et al. (2008) examined the factorial structure of measures of working memory and fluency tasks and found support for specific factors corresponding to category fluency, letter fluency, and working memory, but these factors were strongly correlated and were explained by a second-order, global executive factor.

The literature on factor structure of cognitive test batteries consistently supports the presence of factors involving verbal abilities, spatial abilities, and memory/learning. Factors accounting for measures of attention, speed of processing, working memory, and executive tasks like inhibition and set shifting have been somewhat less consistent, but at least some of this inconsistency relates to the differences in the groups of measures included in the different studies. It is axiomatic that factor structure is dependent upon the specific measures included in the factor analysis. Executive function is a relatively new focus in cognitive psychology and neuropsychology, and consequently, many earlier studies did not comprehensively represent measures of executive abilities.

The current study used confirmatory factor analysis to systematically test how well alternative, a priori models account for associations among NIHTB-CHB and Gold Standard tests. The alternative models ranged from a simple 1-factor model representing global cognition to a 6-factor model corresponding to the six NIHTB-CHB sub-domains. These models also included a 2-factor model representing crystallized and fluid abilities, and models with different levels of differentiation of speed of processing, executive function, and working memory. It was hypothesized that the 6-factor model underlying development of the NIHTB-CHB would provide the best fit, and that NIHTB-CHB tests and their gold standard counterparts would define the same factors.

METHOD

Participants

The participants in this study are described in detail in the Introduction to this series. Briefly, the sample included 268 adults ranging in age from 20 to 85; by design, we did not sample ages 61–64. There were 149 females and 119 males; 148 were non-Hispanic whites, 75 were African Americans, 38 were Hispanics, and 7 were identified as multiracial. Mean age (*SD*) was 52.3 (21.0) years, and mean education (*SD*) was 13.4 (2.9) years. Education was further categorized as less than high school graduate (25%), high school graduate or some college (37%), and Bachelor's degree or higher (38%). Table 1 shows the sample demographics. Data was collected at multiple sites under research protocols approved by site institutional review boards.

https://doi.org/10.1017/S1355617714000307 Published online by Cambridge University Press

Table 1. Adult validation sample demographics

		Gender		Race/ethnicity		
Age groups	Education	Male	Female	White	Black	Hispanic/ Other
20–60 yrs.	< High school	22	26	21	15	12
$N = 159$	High school graduate	29	31	26	19	15
	College +	24	27	24	15	12
65–85 yrs.	< High school	9	11	9	10	1
$N = 109$	High school graduate	12	27	26	11	2
	College +	23	27	42	5	3
TOTAL $N = 268$		119	149	148	75	45

Note. Domains refer to alternative factor models are listed in order from most specific to most general. Toolbox measures are bolded.

Measures

NIHTB-CHB and Gold Standard tests are listed in Table 2. Development of NIHTB-CHB tests is described in detail in individual articles in this series, and Gold Standard tests also are described in more detail in individual articles.

NIHTB-CHB Measures

Seven tests from the NIHTB-CHB were used. These included the Dimensional Change Card Sort (DCCS) Test (executive function domain), the Flanker Inhibitory Control Test (executive function), the Picture Sequence Memory Test (episodic memory), the Picture Vocabulary Test (vocabulary), the Oral Reading Recognition Test (reading), the List Sorting Working Memory Test (working memory), and the Pattern Comparison Processing Speed Test (processing speed).

The NIHTB-CHB included measures of two important components of executive function, flexibility/set shifting (DCCS) and inhibitory control (Flanker Incongruent). The DCCS is a measure of cognitive flexibility. Two target pictures are presented that vary along two dimensions (e.g., shape and color). Participants are asked to match a series of bivalent test pictures (e.g., yellow balls and blue trucks) to the target pictures, first according to one dimension (e.g., color) and then, after several trials, according to the other dimension (e.g., shape). "Switch" trials are also used, in which the participant must change the dimension being matched, thus requiring the cognitive flexibility to quickly choose the correct stimulus. The Flanker task measures attention and inhibitory control. The test requires the participant to focus on a given stimulus while inhibiting attention to stimuli (fish for ages 3–7 or arrows for ages 8–85) flanking it. Sometimes

Table 2. Measures and associated domains/dimensions

Measure	Associated domains
Vocabulary	Vocabulary, Language, Crystalized/Global
Reading	Reading, Language, Crystalized, Global
Picture Sequence Memory	Episodic Memory, Fluid, Global
List Sorting	Working Memory, Executive, Fluid, Global
Flanker Incongruent	Executive, Fluid, Global
DCCS	Executive, Fluid, Global
Pattern Comparison	Speed, Executive, Fluid, Global
PPVT-R	Vocabulary, Language, Crystalized/Global
WRAT-IV	Reading, Language, Crystalized, Global
RVLT	Episodic Memory, Fluid, Global
BVMT-R	Episodic Memory, Fluid, Global
PASAT	Working Memory, Executive, Fluid, Global
Wechsler Letter Number Sequencing	Working Memory, Executive, Fluid, Global
Wechsler Digit Symbol/Coding	Speed, Executive, Fluid, Global
Wechsler Symbol Search	Speed, Executive, Fluid, Global
WCST Total Errors	Executive, Fluid, Global
DKEFS Stroop Interference	Executive, Fluid, Global

Note. Domains refer to alternative factor models and are listed in order from most specific to most general. Toolbox measures are bolded. DCCS = Dimensional Change Card Sort; PPVT-R = Peabody Picture Vocabulary Test – Revised; WRAT-IV = Wide Range Reading Test – Fourth Edition; RVLT = Rey Auditory Verbal Learning Test; BVMT-R = Brief Visuospatial Memory Test-Revised; PASAT = Paced Auditory Serial Attention Test; WCST = Wisconsin Card Sorting Test.

https://doi.org/10.1017/S1355617714000307 Published online by Cambridge University Press

the middle stimulus is pointing in the same direction as the "flankers" (congruent) and sometimes in the opposite direction (incongruent). Scoring for both the DCCS and Flanker incongruent is based on an algorithm that combines reaction time and accuracy, but for adults where accuracy is generally high, is substantially based on reaction time.

Processing speed is measured with the Pattern Comparison task. Participants are asked to discern whether two side-by-side pictures are the same or not. The items are designed to be simple and easily dicriminable to most purely measure processing speed. The participants' raw score is the number of items correct in a 90-s period.

The List Sorting test measures working memory. This test requires immediate recall and sequencing of different visually and orally presented stimuli. Pictures of different foods and animals are displayed with accompanying audio recording and written text (e.g., "elephant"), and the participant is asked to say the items back in size order from smallest to largest, first within a single dimension (either animals or foods, called 1-List) and then on two dimensions (foods, then animals, called 2-List). The score is equal to the number of items recalled and sequenced correctly.

The NIHTB-CHD assesses episodic memory using the Picture Sequence Memory Test. It involves recalling series of illustrated objects and activities that are presented in a particular order on the computer screen. The participants are asked to recall the sequence of pictures that is demonstrated over two learning trials; sequence length varies from 6 to 18 pictures, depending on age. Participants are given credit for each adjacent pair of pictures (i.e., if pictures in locations 7 and 8 and placed in that order and adjacent to each other anywhere—such as slots 1 and 2—one point is awarded) they correctly place, up to the maximum value for the sequence, which is one less than the sequence length (if there are 18 pictures in the sequence, the maximum score is 17, because that is the number of adjacent pairs of pictures that exist).

Language is assessed using the Picture Vocabulary Test. This measure of receptive vocabulary is administered in a computerized adaptive format. The participant is presented with an audio recording of a word and four photographic images on the computer screen and is asked to select the picture that most closely matches the meaning of the word. Reading is measured in a computerized adaptive format with the Oral Reading Recognition Test. The participant is asked to read and pronounce letters and words as accurately as possible. For both Vocabulary and Reading, an item response theory ability score is calculated based on the specific items administered.

Gold standard measures

"Gold Standard" cognitive measures included the Reading subtest from the Wide Range Achievement Test – Fourth Edition Reading subtest (Wilkinson & Robertson, 2006) (reading), the Peabody Picture Vocabulary Test – Fourth Edition (Dunn & Dunn, 2007)(vocabulary), the Wechsler Adult Intelligence Scale – Fourth Edition Letter-Number Sequencing (working memory), Coding/Digit Symbol (processing speed), and Symbol Search (processing speed) subtests (Wechsler, 2008), the Delis-Kaplan Executive Function System (Delis, Kramer, & Kaplan, 2001) Color-Word Interference score (executive function), the total learning score from the Brief Visuospatial Memory Test – Revised (Benedict, 1997) (episodic memory), the total learning score from the Rey Auditory Verbal Learning Test (Rey, 1964) (episodic memory), and the Paced Auditory Serial Addition Test (Gronwall, 1977; first channel only)(working memory), and the Wisconsin Card Sorting Test – total errors (Heaton, Chelune, Talley, Kay, & Curtiss, 1993) (executive function).

Table 3. Alternate dimensional models underlying NIHTC-CTB and gold standard measures

1f – Global Cognition
2f – Crystalized, Fluid
2f – Memory, Non-Memory
3f – Language, Memory/Working Memory, Executive/Speed
3f – Language, Memory, Working Memory/Executive/Speed
4f – Language, Memory, Working Memory, Executive/Speed
4f – Vocabulary, Reading, Memory, Working Memory/Executive/Speed
4f – Vocabulary, Reading, Memory/Working Memory, Executive/Speed
5f – Language, Memory, Working Memory, Executive, Speed
5f – Vocabulary, Reading, Memory, Working Memory, Executive/Speed
6f – Vocabulary, Reading, Memory, Working Memory, Executive, Speed

Data Analysis

Latent variable modeling methods were used to test convergent and discriminant validity of NIHTB-CHB and Gold Standard measures. The basic process was to perform a series of confirmatory factor analyses to test alternative models for the dimensions hypothesized to underlie the NIHTB-CHB and Gold Standard tests. The alternative models that were tested are shown in Table 3. These models all included parameters to estimate residual covariances of measures that shared a common method (WAIS-R Digit Symbol and Symbol Search which both are counts of number of correct responses in a specific time period, DCCS and Flanker which both are computer presented and are strongly based on reaction time). The models shown in Table 3 were separately estimated and model fit indices were compared to identify the best fitting model. Model fit and model parsimony were considered in identifying the best model. Extremely high correlations among latent factors ($\geq$0.90) were considered to provide evidence in favor a more parsimonious, lower dimensional solution. The best fitting model at this stage had a simple structure with each indicator loading on just one factor. Modification indices were then examined to identify cross loadings of NIHTB-CHB measures on other factors that would significantly improve model fit if freely estimated. Convergent validity for a NIHTB-CHB measure was evidenced by a strong loading on the dimension corresponding to the primary conceptual domain. Discriminant validity was shown if no loading, or a smaller loading, was required for a NIHTB-CHB measure on a secondary dimension/domain.

A secondary analysis using multiple group confirmatory factor analysis tested invariance of the best model across two adult age groups, 20- to 60-year-olds and 65- to 85-year-olds. In multiple group CFA, a common model for both groups is specified on an a priori basis, and then group differences in individual parameters can be systematically tested. The best fitting model from the previous stage of analysis was used as the starting point. A multiple group model was fitted with loadings and intercepts that were constrained to be equal in the two groups, but common factor means, variances, and covariances, and residual variances for individual indicators were allowed to differ across groups. A second, freely estimated model allowed loadings and intercepts to vary across groups. One loading and one intercept for each factor were constrained to equality to identify this second multiple group model. Improvement in model fit associated with freely estimating loadings and intercepts in the two groups was evaluated using the change in the comparative fit index (CFI; Bentler, 1990) as recommended by Cheung and Rensvold (2002), and specifically, a difference in the CFI values greater than 0.01 was used as the standard for identifying significant measurement non-invariance in the two groups.

Variables were recoded before analysis using the Blom rank order normalization algorithm in SAS Proc Rank. This resulted in variables with relatively normal distributions and also established a common scale of measurement of all variables. Scores for DKEFS Interference and WCST Errors were inverted so that higher scores indicated better performance on all measures. Normalized scores were multiplied by 3.0 and added to 10.0 to place them on a common scale with mean of 10.0 and standard deviation of 3.0.

Model estimation was performed with Mplus version 7.0 (Muthén & Muthén, 1998–2012) using a maximum likelihood estimator for continuous variables applied to a mean and covariance data structure. Latent variable modeling traditionally uses an overall chi square test of model fit, often supplemented by several fit indices to better characterize model fit. Commonly used fit indices include the CFI, the Tucker-Lewis index (TLI; Tucker & Lewis, 1973), the root mean square error of approximation (RMSEA; Browne & Cudek, 1993), and the standardized root mean squared residual (SRMR; Bentler, 1995). The χ^2 difference test (Steiger, Shapiro, & Browne, 1985) was used to determine if fit significantly improved as a result of freeing one or more parameters in a model. Modification indices correspond to the improvement in model fit as measured by the amount the overall χ^2 value would decrease if a constrained parameter were freely estimated. A threshold of 6.63 was used as a standard for significant improvement in fit, which corresponds to $p = .01$ for a χ^2 variate with 1 degree of freedom.

RESULTS

The 5-factor model (Vocabulary, Reading, Episodic Memory, Working Memory, Executive/Speed) and 6-factor model (Vocabulary, Reading, Episodic Memory, Working Memory, Executive, Speed) both showed good fit and were very

Table 4. Fit indices for alternate models of cognitive dimensions in combined 20- to 60-year-old and 65- to 85-year- old age groups.

Model	Overall χ^2[df]	CFI	TLI	RMSEA (90% CI)	SRMR
1f – global	1076.6 [117]	0.666	0.611	0.175 (0.166–0.185)	0.129
2f – cryst, fluid	429.2 [116]	0.891	0.872	0.101 (0.090–0.111)	0.072
2f – mem, non-mem	1000.7 [116]	0.692	0.638	0.169 (0.159–0.179)	0.126
3f – lang, mem/wm, exec/speed	386.8 [114]	0.905	0.887	0.095 (0.084–0.105)	0.069
3f – lang, mem, wm/exec/speed	362.6 [114]	0.913	0.897	0.090 (0.080–0.101)	0.067
4f – lang, mem, wm, exec/speed	323.9 [111]	0.926	0.909	0.085 (0.074–0.096)	0.059
4f – voc, read, mem, wm/exec/speed	274.6 [111]	0.943	0.930	0.074 (0.063–0.085)	0.060
4f – voc, read, mem/wm, exec/speed	299.0 [111]	0.934	0.920	0.080 (0.069–0.091)	0.063
5f – lang, mem, wm, exec, speed	315.7 [107]	0.927	0.908	0.085 (0.075–0.096)	0.058
5f – voc, read, mem, wm, exec/speed	229.1 [107]	0.957	0.946	0.065 (0.054–0.077)	0.050
6f – voc, read, mem, wm, exec, speed	219.8 [102]	0.959	0.945	0.066 (0.054–0.078)	0.049

All model included residual correlations of measures sharing a common method.

https://doi.org/10.1017/S1355617714000307 Published online by Cambridge University Press

Table 5. Standardized factor loadings (standard errors in parentheses) for 5-factor model. NIHTB-CTB measures are bolded.

Latent factor	Observed indicator	Loading
Reading	**Reading**	**0.969 (0.014)**
	WRAT-R	0.893 (0.017)
Vocabulary	**Vocabulary**	**0.908 (0.020)**
	PPVT-R	0.857 (0.023)
Episodic Memory	**Picture Sequence Memory**	**0.824 (0.026)**
	RAVLT	0.769 (0.031)
	BVMT	0.814 (0.027)
	List Sorting	**0.368 (0.089)**
Working Memory	**List Sorting**	**0.448 (0.087)**
	PASAT	0.692 (0.038)
	Wechsler Letter Number Sorting	0.780 (0.032)
Executive/Speed	**Flanker**	**0.711 (0.037)**
	DCCS	**0.741 (0.034)**
	Wisconsin Card Sort Total Errors	0.626 (0.042)
	DKEFS Stroop Interference	0.800 (0.027)
	Pattern Comparison	**0.644 (0.040)**
	Wechsler Digit Symbol	0.771 (0.030)
	Wechsler Symbol Search	0.790 (0.028)

Table 6. Intercorrelation of factors from 5-factor model

	Reading	Vocabulary	Episodic Memory	Working Memory
Vocabulary	0.820			
Episodic Memory	0.295	0.135		
Working Memory	0.579	0.481	0.771	
Executive	0.389	0.213	0.808	0.871

similar in terms of overall model fit (See Table 4). However, there was a technical problem with the 6-factor model such that the estimated correlation of the Speed and Executive factors exceeded 1.0 (1.013). In addition, when fit of these two non-nested models was compared using Information Criteria, the 5-factor model showed slightly better fit, that is, a smaller value (Akaike's Information Criterion, 5-factor – 19219.4, 6-factor – 19220.0; Bayesian Information Criterion, 5-factor – 19445.4, 6-factor – 19464.0; Sample-Size Adjusted Bayesian Information Criterion, 5-factor – 19245.6, 6-factor – 19248.4). Consequently, the 5-factor model with executive and speed combined was selected as the best model. Allowing List Sorting to cross-load on the Episodic Memory factor significantly improved model fit ($\chi^2[1] = 12.6$; $p < .001$) but no other cross-loadings were identified. Thus, the final model included the five *a-priori* specified factors described in Table 3 along with three additional parameters: a residual correlation of WAIS-R Digit Symbol with WAIS-R Symbol Search, a residual correlation of DCCS with Flanker, and a loading of List Sorting on the Episodic Memory factor.

Standardized loadings for the 5-factor model are presented in Table 5. Loadings for the NIHTB-CHB variables Reading and Vocabulary on their respective factors exceeded 0.90. Picture Sequence Memory had a standardized loading of 0.82 on the Episodic Memory factor. DCCS and Flanker had loadings of 0.70–0.75 on the Executive/Speed factor, and Pattern Comparison had a loading on this factor of about 0.65 on Executive/Speed. List Sorting had a loading of 0.45 on the Working Memory factor and a secondary loading of 0.37 on Episodic Memory. Overall, these findings show strong evidence of convergent validity. The presence of only one, relatively weak, cross loading supports discriminant validity of the NIHTB-CHB. While loadings of NIHTB-CHB measures of executive function and processing speed on the Executive/Speed factor were strong, these convergent validity estimates were weaker than for other NIHTB-CHB measures. This is not surprising because of the relative heterogeneity of the indicators for these factors and the absence of direct gold standard analogues of the Toolbox measures such as were available for Vocabulary and Flanker does not have a direct gold standard analogue, and while DCCS and WCST both assess flexibility and set shifting, DCCS is strongly based on reaction time while WCST is essentially an untimed measure of accuracy.

The intercorrelations of the five factors ranged from 0.14 to 0.87 (See Table 6). The Working Memory factor was highly correlated with Executive/Speed, and to slightly lesser extent, Episodic Memory. Executive/Speed and Episodic Memory were highly correlated. Vocabulary and Reading similarly were highly correlated.

The pattern of intercorrelations of the five factors suggested that higher order factors representing crystallized (Vocabulary, Reading) and fluid abilities (Episodic Memory, Executive/Speed, Working Memory) might explain these correlations. A secondary analysis fitted a hierarchical model adding second order factors for Crystallized and Fluid Abilities to explain the first order factors. There were technical difficulties with estimating this hierarchical model using the maximum likelihood estimator; specifically with estimating the loading of the Reading factor on the second order Crystallized Abilities factor. Consequently, a Bayesian estimator (Muthén & Muthén, 1998–2012) was used, and this model converged and was interpretable. Vocabulary ($\lambda = 0.84$) and Reading ($\lambda = 0.99$) had strong loading on the Crystallized Abilities factor, and Episodic Memory ($\lambda = 0.85$), Executive/Speed ($\lambda = 0.93$), and Working Memory ($\lambda = 0.95$) had strong loadings on the Fluid Abilities factor. Crystallized Abilities and Fluid Abilities were moderately correlated ($r = 0.46$).

Multiple group CFA was used to test invariance of the 5-factor model across two age groups (20–60 years, $n = 159$; 65–85 years, $n = 109$). The initial multiple group model constrained loading and intercepts to be the same in the two groups, and resulted in good model fit ($\chi^2[239] = 372.4$; $p < .001$; CFI = 0.945; TLI = 0.937; RMSEA = 0.065 (0.052–0.077); SRMR = 0.083). Loadings and intercepts and the cross loading of List Sorting on the Episodic Memory factor were then freely estimated in the two groups; model fit ($\chi^2[214] = 324.1$; $p < .001$; CFI = 0.954; TLI = 0.942,

https://doi.org/10.1017/S1355617714000307 Published online by Cambridge University Press

RMSEA = 0.062 (0.048–0.075); SRMR = 0.062) was not significantly better using the change in CFI criterion recommended by Cheung and Rensvold (2002). These results support factorial invariance of the 5-factor model across the 20- to 60- and 65- to 85-year age groups.

DISCUSSION

NIHTB-CHB measures and their Gold Standard analogues consistently defined factors in a pattern that supported the convergent and discriminant validity of the NIHTB-CHB. Results showed that the empirical dimensions underlying the NIHTB-CHB and Gold Standard tests corresponded to the guiding conceptual model. Tests measuring the same ability had strong loadings on the same factor. Only one cross-loading on a not hypothesized secondary factor was identified (List Sorting on Episodic Memory). This secondary loading is conceptually plausible and was smaller than the primary loading of List Sorting on Working Memory.

A 5-factor solution was identified as the best model. The 6-factor solution based on the conceptual model that guided the development of the NIHTB-CHB had very similar overall fit, but measures of executive function and processing speed were not clearly separable. The fact that the NIHTB-CHB executive measures are based substantially on speed of executive operations probably accounts for this. The only difference between the 5-and 6-factor models was that speed and executive measures defined one factor in the former and two in the latter. Previous literature has identified processing speed as an important if not integral component of executive function (Albinet, Boucard, Bouquet, & Audiffren, 2012; Salthouse, 2005). The results of this study provide further evidence that these cognitive processes are integrally related. When executive function and speed were modeled as separate factors, their correlation was very high even though NIHTB-CHB executive measures (Flanker and DCCS) were defined both by speed and accuracy and WCST Errors was not a timed measure. From a practical perspective, the speed and executive measures in the NIHTB-CHB are not likely to provide different information in future studies. An additional implication is that the NIHTB-CHB measures will have limitations for differentiating speed and executive components of cognitive abilities.

Episodic Memory was identified as a clearly separable factor and this is consistent with several studies based on different measures of cognition that have identified one or more episodic memory factors (Bowden, Carstairs, & Shores, 1999; Mungas et al., 2011; Smith et al., 1992, 1993; Tuokko et al., 2009). Reading and Vocabulary also represented separable dimensions in this study. Since previous studies have shown a more general language factor, for example within the Wechsler Adult Intelligence Scale (Smith et al., 1992, 1993), it would not have been surprising to find that vocabulary and reading share a common factor. In fact, these two factors were highly correlated ($r = 0.82$), but model fit was better when they were separated as opposed to combined in a common language factor. The Working Memory factor identified in this study is also noteworthy in that it was separable from both episodic memory and executive/speed, but was highly correlated with these factors consistent with expectations based on shared anatomical substrates and cognitive processes. That is, working memory is often considered a component of executive function based on shared anatomy involving frontal/subcortical brain systems. Attention and working memory are prerequisites for episodic memory processes involved in placing information into short- and long-term memory stores that represent learning and memory, and so, are parts of an integrated cognitive system underlying episodic memory.

Results of this study showed poor fit for a 1-factor model representing global cognition. Fit was substantially better for a 2-factor model representing crystallized and fluid abilities, but still was not at a level that would indicate good, or even adequate fit to the data. However, correlations among the five identified factors in the final model were substantial and did support the general distinction between fluid and crystallized abilities. Language and Reading were much more highly correlated with one another ($r = 0.82$) than with Episodic Memory, Executive/Speed, and Working Memory (r's ranging from 0.14 to 0.58), and similarly, correlations among the latter three factors (r's ranging from 0.77 to 0.87) exceeded their correlations with Reading and Vocabulary. The hierarchical factor model supported the presence of second order Crystallized and Fluid Abilities factors. The practical implications of these results are twofold. First, this study supports a relatively fine-grained characterization of cognition into five correlated but differentiable dimensions that might be used in future studies involving cognition. Second, it probably is reasonable to combine measures into crystallized and fluid composites if less refined differentiation of cognition is required to meet study goals. This would have advantages associated with simpler study design, analysis, and interpretation, and with higher reliability of cognitive measures as a result of having more items contribute to the composite measures.

An important goal of the NIHTB was to develop measures that can be used for longitudinal studies across the age span from 3 to the upper end of the adult age span. The five-factor solution identified in this study was the same as that identified in a previous publication that examined the factor structure of these measures in children in the 8–15 years age range (Mungas et al., 2013). That study formally tested and strongly supported invariance across 8- to 15-year-olds and adults on the same 5-factor solution that was identified in this study. This study extended the evaluation of invariance across the adult age span and showed factorial invariance across 20- to 60-year-olds and 65- to 85-year-olds. Results of these two studies collectively show that factors accounting for NIHTB-CHB tests are remarkably consistent across age groups starting at 8 years. This suggests that it will be possible to use the NIHTB-CHD to measure cognition in a comparable way from age 8 to 85. Measurement invariance at different time points is a prerequisite for longitudinal studies, and this presents a special challenge for studies that extend across qualitatively and

https://doi.org/10.1017/S1355617714000307 Published online by Cambridge University Press

quantitatively different developmental stages of the life span. Results of this study show evidence that cognition is similarly structured from age 8 years into late adulthood along with evidence that specific NIHTB-CHB measures relate to underlying dimensions of cognition in the same way in different age groups. This provides preliminary evidence that the NIHTB-CHB will be useful for longitudinal studies across the lifespan.

There are important limitations of this study. First, the sample size is relatively small for testing factorial invariance across different age groups and replication of these results with different samples is important. Second, the sample for this study did not include 61- to 64-year-olds and this limits generalizability of results to this age range. Third, the executive measures in the NIHTB-CHB are reaction time based and this likely is an important factor accounting for the extremely high correlation of speed and executive factors. Consequently, these results may be specific to the measures included in this study and might overestimate relations of speed and executive function in a broader context. Finally, the use of modification indices to evaluate discriminant validity is a limitation. Modification indices are data driven and are subject to capitalization on chance variation. While use of modification indices in this study was limited this nevertheless raises a concern about replicability of the findings related to discriminant validity, and it will be important for future studies to evaluate discriminant validity in different samples.

The use of factor analysis to establish construct validity has been questioned by Delis, Jacobson, Bondi, Hamilton, and Salmon (2003), and more specifically, they argued that factor structures found in normative samples might not apply in clinical populations. Bowden (2004) identified methodological concerns with the analyses Delis et al. used to support their arguments, but nevertheless, construct validity is a cumulative scientific process and this study is first step toward characterizing the construct validity of the NIHTB-CTD. This battery was not designed for use with clinical populations, but future research to directly test measurement invariance across normative and clinical samples could be very important for defining the range of utility of this test battery.

The NIHTB-CHB is a battery of cognitive tests developed using state of the art measurement and administration methods and technology. Results of this study broadly support the construct validity of this test battery in relation to the formal conceptual model that guided its development. This is an important, but also early step in the ongoing validation of the NIHTB-CHD. Future research will be required to define how NIHTB-CHB measures and change in these measures relate to relevant brain and non-brain criteria. Results of this study are limited to the English-language version of the tests, and validation with Spanish speakers is an important future goal. The careful development process of the NIHTB leading to broad availability will likely promote widespread use of these measures, and this aggregated data will support studies to further define the validity and utility of the NIHTB-CHD in a variety of specific contexts.

ACKNOWLEDGMENTS

This study is funded in whole or in part with Federal funds from the Blueprint for Neuroscience Research, NIH, under Contract No. HHS-N-260-2006-00007-C. The authors have no financial conflicts of interest related to this work.

REFERENCES

Albinet, C.T., Boucard, G., Bouquet, C.A., & Audiffren, M. (2012). Processing speed and executive functions in cognitive aging: How to disentangle their mutual relationship? *Brain and Cognition*, *79*(1), 1–11.

Benedict, R. (1997). *Brief Visuospatial Memory Test-Revised professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Bentler, P.M. (1990). Comparative fit indices in structural models. *Psychological Bulletin*, *107*, 238–246.

Bentler, P.M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Bowden, S.C. (2004). The role of factor analysis in construct validity: Is it a myth? *Journal of the International Neuropsychological Society*, *10*(7), 1018–1019.

Bowden, S.C., Carstairs, J.R., & Shores, E.A. (1999). Confirmatory factor analysis of combined Wechsler Adult Intelligence Scale-Revised and Wechsler Memory Scale-Revised scores in a healthy community sample. *Psychological Assessment*, *11*(3), 339–344.

Bowden, S.C., Cook, M.J., Bardenhagen, F.J., Shores, E.A., & Carstairs, J.R. (2004). Measurement invariance of core cognitive abilities in heterogeneous neurological and community samples. *Intelligence*, *32*(4), 363–389.

Browne, M., & Cudek, R. (1993). Alternate ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 136–162). Thousand Oaks, CA: Sage.

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255.

Crane, P.K., Narasimhalu, K., Gibbons, L.E., Pedraza, O., Mehta, K. M., Tang, Y., … Mungas, D.M. (2008). Composite scores for executive function items: Demographic heterogeneity and relationships with quantitative magnetic resonance imaging. *Journal of the International Neuropsychological Society*, *14*(5), 746–759.

Delis, D.C., Jacobson, M., Bondi, M.W., Hamilton, J.M., & Salmon, D.P. (2003). The myth of testing construct validity using factor analysis or correlations with normal or mixed clinical populations: Lessons from memory assessment. *Journal of the International Neuropsychological Society*, *9*, 936–946.

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

Dunn, L.M., & Dunn, D.M. (2007). *Peabody Picture Vocabulary Test-Fourth Edition (PPVT-4)*. Circle Pines, MN: American Guidance Services.

Gershon, R.C., Wagster, M.V., Hendrie, H.C., Fox, N.A., Cook, K.F., & Nowinski, C.J. (2013). NIH toolbox for assessment of neurological and behavioral function. *Neurology*, *80*(11 Suppl 3), S2–S6.

Gronwall, D.M. (1977). Paced auditory serial-addition task: A measure of recovery from concussion. *Perceptual and motor skills*, *44*, 367–373.

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

Larrabee, G.J., Kane, R.L., & Schuck, J.R. (1983). Factor analysis of the WAIS and Wechsler Memory Scale: An analysis of the construct validity of the Wechsler Memory Scale. *Journal of Clinical Neuropsychology*, *5*(2), 159–168.

Miyake, A., Friedman, N.P., Emerson, M.J., Witzki, A.H., Howerter, A., & Wager, T.D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

Mungas, D., Widaman, K., Zelazo, P.D., Tulsky, D., Heaton, R.K., Slotkin, J., … Gershon, R.C. (2013). NIH toolbox cognitive health battery (CB): Factor structure for 3- to 15 year olds. *Monographs Society for Research on Child Development*, *78*(4), 103–118. Chapter VII.

Mungas, D., Widaman, K.F., Reed, B.R., & Tomaszewski Farias, S. (2011). Measurement invariance of neuropsychological tests in diverse older persons. *Neuropsychology*, *25*(2), 260–269.

Muthén, L.K., & Muthén, B.O. (1998–2012). *Mplus User's Guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Rey, A. (1964). *L'examen clinique en psychologie*. Paris: Presses Universitaires de France.

Salthouse, T.A. (2005). Relations between cognitive abilities and measures of executive functioning. *Neuropsychology*, *19*(4), 532–545.

Smith, G.E., Ivnik, R.J., Malec, J.F., Kokmen, E., Tangalos, E.G., & Kurland, L.T. (1992). Mayo's Older Americans Normative Studies (MOANS): Factor structure of a core battery. *Psychological Assessment*, *4*(3), 382–390.

Smith, G.E., Ivnik, R.J., Malec, J.F., & Tangalos, E.G. (1993). Factor structure of the Mayo Older Americans Normative Sample (MOANS) core battery: Replication in a clinical sample. *Psychological Assessment*, *5*(1), 121–124.

Steiger, J.H., Shapiro, A., & Browne, M.W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika*, *50*, 253–264.

Tucker, L.R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10.

Tuokko, H.A., Chou, P.H., Bowden, S.C., Simard, M., Ska, B., & Crossley, M. (2009). Partial measurement equivalence of French and English versions of the Canadian Study of Health and Aging neuropsychological battery. *Journal of the International Neuropsychological Society*, *15*(3), 416–425.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale IV*. San Antonio, TX: Harcourt Assessment.

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