Published in final edited form as:
J Int Neuropsychol Soc. 2014 July ; 20(6): 588–598. doi:10.1017/S1355617714000241.

Reliability and Validity of Composite Scores from the NIH Toolbox Cognition Battery in Adults

Robert K. Heaton[1], **Natacha Akshoomoff**[1], **David Tulsky**[2], **Dan Mungas**[3], **Sandra Weintraub**[4], **Sureyya Dikmen**[5], **Jennifer Beaumont**[6], **Kaitlin B. Casaletto**[7], **Kevin Conway**[8], **Jerry Slotkin**[6], and **Richard Gershon**[6]

[1]Department of Psychiatry, University of California, San Diego, California [2]Departments of Rehabilitation Medicine, Orthopedic Surgery, and General Medicine, New York University Langone Medical Center, New York; and Spinal Cord Injury, Kessler Foundation, West Orange, New Jersey [3]Department of Neurology, University of California, Davis, California [4]Cognitive Neurology and Alzheimer's Disease Center and Department of Psychiatry and Behavioral Sciences, Northwestern Feinberg School of Medicine, Chicago, Illinois [5]Department of Rehabilitation Medicine, University of Washington, Seattle, Washington [6]Department of Medical Social Sciences, Northwestern University, Chicago, Illinois [7]San Diego State University/University of California, San Diego Joint Doctoral Program in Clinical Psychology, San Diego, California [8]National Institute on Drug Abuse, Rockville, Maryland

Abstract

This study describes psychometric properties of the NIH Toolbox Cognition Battery (NIHTB-CB) Composite Scores in an adult sample. The NIHTB-CB was designed for use in epidemiologic studies and clinical trials for ages 3 to 85. A total of 268 self-described healthy adults were recruited at four university-based sites, using stratified sampling guidelines to target demographic variability for age (20–85 years), gender, education, and ethnicity. The NIHTB-CB contains seven computer-based instruments assessing five cognitive sub-domains: Language, Executive Function, Episodic Memory, Processing Speed, and Working Memory. Participants completed the NIHTB-CB, corresponding gold standard validation measures selected to tap the same cognitive abilities, and sociodemographic questionnaires. Three Composite Scores were derived for both the NIHTB-CB and gold standard batteries: "Crystallized Cognition Composite," "Fluid Cognition Composite," and "Total Cognition Composite" scores. NIHTB Composite Scores showed acceptable internal consistency (Cronbach's alphas = 0.84 Crystallized, 0.83 Fluid, 0.77 Total), excellent test–retest reliability (r: 0.86–0.92), strong convergent (r: 0.78–0.90) and discriminant (r: 0.19–0.39) validities *versus* gold standard composites, and expected age effects ($r = 0.18$ crystallized, $r = -0.68$ fluid, $r = -0.26$ total). Significant relationships with self-reported prior school difficulties and current health status, employment, and presence of a disability provided evidence of external validity. The NIH Toolbox Cognition Battery Composite Scores have

Correspondence and reprint requests to: Robert K. Heaton, Department of Psychiatry, UCSD School of Medicine, 9500 Gilman Drive, La Jolla, CA 92093-0603. rheaton@ucsd.edu.

excellent reliability and validity, suggesting they can be used effectively in epidemiologic and clinical studies.

Keywords

Neuropsychological assessment; Memory; Executive Function; Attention; Cognitive assessment; Cognitive screener

INTRODUCTION

The NIH Toolbox Cognition Battery (NIHTB-CB), together with test modules for motor, sensory, and emotional functioning, comprise the "NIH Toolbox for the Assessment of Neurological and Behavioral Function." The development of the NIH Toolbox was commissioned by 16 NIH institutes to provide brief, efficient and highly accessible measures for broad use in future epidemiologic and clinical research. Additional important goals of the NIH Toolbox initiative were to use nonproprietary instruments that could be administered in both English and Spanish, and that would be able to tap behavioral constructs across the lifespan (ages 3 to 85 years) (Weintrab *et al.*, 2013).

The NIHTB-CB is currently composed of seven test instruments that measure abilities within 5 major cognitive domains. The individual test instruments, described in detail in other papers within this series, include the Dimensional Change Card Sort Test (DCCS) (Executive Function-Cognitive Flexibility) and the Flanker Inhibitory Control and Attention Test (Executive Function-Inhibitory Control and Attention); the Picture Sequence Memory Test (Episodic Memory); the Picture Vocabulary Test (Language-Vocabulary Comprehension); the Oral Reading Recognition Test (Language-Reading Decoding); the List Sorting Working Memory Test (Working Memory); and the Pattern Comparison Processing Speed Test (Processing Speed).

Some researchers will want to consider measures of these various cognitive functions separately, but others are likely to focus on a smaller number of composite scores that represent overall cognition and/or certain categories of abilities. Such composite scores can be defined using factor analytic methods (see Mungas et al., 2013, and Mungas et al., this issue), but these yield different combinations of scores for different age groups and consequently may not be well suited to longitudinal research or research that spans multiple age ranges (e.g., early childhood to later adulthood).

Another approach to defining composite scores is to group tests that may tap more than one specific ability domain but share certain theoretical and psychometric characteristics across the lifespan (Akshoomoff et al., 2013). In the two-component theory of intellectual development (Cattell, 1971; Horn, 1968, 1970), for example, the premise is that the organization of *fluid* and *crystallized* abilities develops and transforms throughout the life span (Li et al., 2004). Fluid abilities are used to solve problems, think and act quickly, and encode new episodic memories, and play an important role in adapting to novel situations in everyday life. Fluid abilities are presumably relatively dynamic, based on "online" and in real-time processes, and are less dependent on past learning experiences and cultural biases

than on biological processes that affect current brain function. Crystallized abilities, in contrast, represent an accumulated store of verbal knowledge and skills, and are thus more heavily influenced by education and cultural exposure. Crystallized abilities show rapid developmental change during childhood, typically continue to improve slightly into middle adulthood, and then remain relatively stable in old age. In contrast, fluid abilities improve rapidly in childhood and typically peak in early adulthood. However, they then tend to be more sensitive to neurobiological integrity, including changes in brain structure and function associated with aging and a variety of neurological disorders such as acquired dysfunction due to traumatic brain injury, stroke, and dementing illnesses.

Here, we present data from the adult validation sample for the NIHTB Cognition Battery that is based on three summary scores: Crystallized Cognition Composite, Fluid Cognition Composite, and Total Cognition Composite (a combination of both crystallized and fluid scores). Results are from 268 self-described healthy adults, ages 20 to 85. We expected the Crystallized Cognition composite score to remain relatively stable throughout this age range, whereas the Fluid Cognition composite score was expected to show much greater decline with age (see Weintraub et al., this issue). We also present psychometric information, such as internal consistency, test–retest reliability and associations with well accepted, but mostly proprietary, instruments that also putatively tap crystallized and fluid abilities (i.e., “gold standard” measures). We predicted that the NIHTB-CB composite scores would show good convergent and discriminant validities with relevant gold standard measures (i.e., those that putatively tap the same *vs*. different cognitive constructs). These hypotheses were partly based upon the expectation that while fluid and crystallized abilities develop rapidly and roughly in parallel during early childhood, they tend to diverge during adulthood with larger age effects on fluid abilities (Horn & Cattell, 1967; Sattler, 2001; Weintraub et al., submitted; WAIS-III WMS-III Technical Manual, 1997).

With both children and adults, it is important to evaluate the potential impact of demographic variables on various neuropsychological tests (Heaton, Ryan, & Grant, 2009; Heaton, Taylor, & Manly, 2003). For example, information about which demographic variables are associated with performance in healthy individuals can inform important group matching decisions in future research, as well as the creation and use of standards for evaluating performance relative to norms. In addition to predicted changes with age, performance on certain measures also is associated with differences in education, family income, gender, and race/ ethnicity (e.g., Heaton et al., 2003; Heaton, Miller, Taylor, & Grant, 2004). The relationship of each of these demographic variables with the composite measures of NIHTB-CB performance was examined.

Finally, to further explore validity of the NIHTB-CB composite measures, we examined associations between the cognitive summary scores and a few relatively gross measures of health and everyday functioning.

METHODS

Participants

All demographic and health status characteristics of the study cohort were gathered *via* self-report questionnaires. The subject sample included 268 adults, primarily recruited through community flyers around four university-based testing sites: 25 at North Shore University Health System in Evanston, IL, 92 at Kessler Foundation Research Center in West Orange, NJ, 67 at University of Washington in Seattle, WA, and 84 at Northwestern University Cognitive Neurology and Alzheimer's Disease Center (CNADC) in Chicago, IL. The flyers advertised for healthy volunteers but no further health screening or exclusions were applied. However, 62 of the older individuals were from the CNADC registry and were known to be cognitively healthy. Stratified sampling guidelines were used to enhance demographic variability, resulting in a final sample with a mean age of 52.3 (*SD* = 21.0) years, and a mean education level of 13.4 (*SD* = 2.9) years, including 119 males and 149 females, with 148 having self-described ethnicity of Caucasian (non-Hispanic White), 75 African-American, 38 Hispanic, and 7 multi-racial (the 7 multiracial participants were excluded from ethnicity comparisons due to the small sample size and greater heterogeneity). See Table 4 below for details regarding cell sizes for various demographic combinations.

Additional demographic and health status variables (all obtained by self-report) were based on the categorical information obtained from each participant. Family income was categorized into five levels (<$20,000 [18%], $20,000 to $39,999 [23%], $40,000 to $74,999 [28%], $75,000 to $99,999 [25%], and ≥$100,000 [6%]). Current health status was categorized as Excellent (64%), Good (26%), or Fair to Poor (9%). Current employment status was categorized as "Employed for wages or self-employed" (44%), "Retired" (31%), "Out of work" (12%), or "Other" (homemaker or student) (13%).

A subgroup of 89 participants (33%) completed a retest 7 to 21 days after initial testing (mean = 15.5 days; *SD* = 4.8) to assess test–retest reliability and "practice effects". Written, informed consent was obtained from all participants *via* a protocol approved by the institutional review boards at each of the above four institutions.

NIH Toolbox Cognition Battery Measures

All seven of the NIHTB-CB tests were included in this study. This resulted in two measures of crystallized abilities (the NIHTB Picture Vocabulary Test and Oral Reading Test), as well as five measures of fluid abilities: the NIHTB Dimensional Change Card Sort (DCCS) Test of Executive Function-Cognitive Flexibility, NIHTB Flanker Test of Executive Function-Inhibitory Control and Attention, NIHTB Picture Sequence Memory Test of Episodic Memory, NIHTB List Sorting Working Memory Test, and NIHTB Pattern Comparison Processing Speed Test. Detailed descriptions of the individual NIHTB-CB tests and the derived scores that reflect the multiple domains of cognitive functioning are provided in Weintraub et al. (submitted for this series) and in other papers in this series that focused specifically on individual NIHTB-CB test measures. Raw scores from the NIHTB-CB measures were converted to normally distributed standard scores (scaled scores) having a mean of 10 and a standard deviation of 3. These standard scores were then averaged to

compute the NIHTB-CB Crystallized Cognition Composite, NIHTB-CB Fluid Cognition Composite, and NIHTB-CB Total Cognition Composite scores.

"Gold Standard" Cognitive Measures

To assess concurrent validity, the NIHTB Cognition Team selected well-established, usually proprietary "Gold Standard" (GS) measures of the same constructs targeted by the NIHTBCB tests. These included the Reading subtest from the Wide Range Achievement Test – 4 (Wilkinson & Robertson, 2006) and the Peabody Picture Vocabulary Test—Fourth Edition (Dunn & Dunn, 2007), which were combined for the Gold Standard Crystallized Composite score (see details below under Analyses). The Gold Standard Fluid Composite score was derived from GS measures of Processing Speed [average of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV) Coding and Symbol Search subtests (Wechsler, 2008)], Executive Function-Inhibitory control [the Delis-Kaplan Executive Function System (Delis, Kramer, & Kaplan, 2001) Color-Word Interference score], Executive Function-Cognitive Flexibility (Wisconsin Card Sorting Test-Total Errors, Heaton et al., 1993), Episodic Memory [an average of total learning scores from the Brief Visuospatial Memory Test – Revised (Benedict, 1997) and the Rey Auditory Verbal Learning Test (Rey, 1964)], and Working Memory [average of the Paced Auditory Serial Addition Test (Gronwall, 1977); first channel only and WAIS-IV Letter-Number Sequencing].

Analyses

Raw scores for the individual NIHTB-CB and respective GS measures were converted to normally distributed standard scores (scaled scores) that were *not* age-corrected. This was accomplished for each measure by initially ranking the raw scores, then transforming the ranks to create a standard normal distribution, and re-scaling the distribution to have a mean of 10 and standard deviation of 3. The three respective composite scores for each test battery (NIHTB-CB and GS battery) were calculated by averaging the normalized scaled scores for the relevant test measures (i.e., two for Crystallized, five for Fluid, and seven for Total Cognition Composites), and then re-distributing the calculated composite scores (as above for scaled scores) and re-scaling so that each composite score was normally distributed and had a mean of 10 and a standard deviation of 3.

Cronbach's alphas were used to assess internal consistencies of the composite scores. Both Pearson correlation coefficients and intraclass correlation coefficients (ICC) were calculated to evaluate test–retest reliability (ICCs are more likely than Pearson r's to be influenced by any substantial differences that may occur between test and re-test, for example due to practice effects). Magnitudes of "practice effects" (effect sizes) were computed as performance at time 2 (retest) minus performance at time 1 (baseline), divided by the standard deviation of performances at time 1 (Cohen et al., 1992), and *t* tests for dependent means were used to test for statistical significance of these effects. Convergent validity was assessed with correlations between the NIHTB-CB measure and the comparable "gold standard" measure of the same construct (Crystallized, Fluid, and Total); evidence of discriminant validity consisted of lower correlations with "gold standard" measures of a *different* cognitive construct (e.g., NIHTB-CB Crystallized *vs*. Gold Standard Fluid).

Pearson correlation coefficients between age and test performance were calculated to assess and compare expected "normal cognitive aging" effects on the NIHTB-CB and Gold Standard Cognition measures during adulthood. Patterns of age effects also were examined by plotting education-corrected standard scores (scaled scores) for nine age subgroups (from 20–24 to 81–85). General linear regression models (GLMs) were then performed to examine other demographic associations with performance, adjusted for age and other relevant covariates. For all GLM models, the data were examined for extreme outliers before analysis. The scaling procedure applied to the scores minimized the impact of any remaining extreme values. Regression diagnostics were additionally examined to confirm model assumptions. Effect sizes are reported as Cohen's d, with cutoffs of .20, .50, and .80, indicating small, medium, and large effects, respectively.

RESULTS

Internal Consistency

For the total subject sample (N = 268) adequate internal consistencies (Cronbach's alphas) were obtained with the NIHTC-CB Crystallized (0.84), Fluid (0.83) and Total (0.77) composites.

Test–Retest Reliability and Practice Effects

For the 89 participants who were retested, excellent test–retest correlations were observed: r = .92, .86, and .90 for NIHTB-CB Crystallized, Fluid, and Total Cognition composite scores, respectively; all df = 87 and p < .0001 (see Table 1 for both ICCs and Pearson correlations). There were also almost identical test–retest correlations with the Gold Standard composite scores (r = .93, .95, and .95 for Gold Standard Crystallized, Fluid, and Total, df = 87, p <.0001). Although the retest sample was one of convenience, their representativeness of the total subject sample is supported by their demographically uncorrected scaled score means and SDs on all NIHTB-CB and gold standard composites, which in each case was virtually identical to the expected mean = 10 (SD = 3) (Table 2).

The NIHTB-CB Crystallized Cognition Composite score evidenced a very small, nonsignificant practice effect over an average 2-week test–retest interval: *mean practice effect* in scaled score units = 0.16, SD = 1.20, t (88) = 1.23, p = .22, d = 0.05 (see Table 2). However, the NIHTB-CB Fluid Cognition Composite score showed a small to medium practice effect of about one scaled score point (*mean* = 1.25, SD = 1.62, t (88) = 7.31, p <. 0001, d = 0.42), and the NIHTB-CB Total Cognition Composite had a small practice effect (*mean* = 0.86, SD = 1.35, t (88) = 6.03, p <.0001, d = 0.29). Comparable, but slightly higher practice effects were noted on the Gold Standard composites: 0.29, p = .02, d = 0.10 for Crystallized; 1.14, p <.0001, d = 0.40 for Fluid; 0.89, p <.0001, d = 0.30 for Total.

Construct Validity

Convergent—Pearson correlations demonstrated that there was good convergent validity for the NIHTB-CB and Gold Standard measures on the Crystallized (r = .90), Fluid (r = . 78), and Total Cognition (r = .89) Composite scores (see Table 3).

Discriminant—Evidence for discriminant validity is provided by substantially lower correlations between NIHTB-CB Crystallized and Gold Standard Fluid Cognition Composite scores (r = .39) and between NIHTB-CB Fluid and Gold Standard Crystallized Cognition Composite scores (r = .19; see Table 3). In addition, the NIHTB-CB Crystallized and Fluid composites showed very modest associations with each other (r = 0.17).

Steiger's Z tests were conducted to statistically compare correlations between TB Crystallized and GS Crystallized, *versus* TB Crystallized and GS Fluid (r = 0.90 *vs.* r = 0.39, $p < .001$), and also those between TB Fluid and GS Fluid *versus* TB Fluid and GS Crystallized (r = 0.78 *vs.* r = 0.19, $p < .001$).

NIHTB-CB Performances by Demographic Subgroups

Table 4 reports the demographic (age/education/gender/ethnicity) cell sizes and demographically *un-corrected* means and SDs for the NIHTB-CB Total Cognition Composite (scaled score metric with overall mean = 10, SD = 3) by demographic categories (i.e., age, education, ethnicity, gender cells). In general, and as expected, better performances tend to be associated with younger age and higher education, as well as with non-Hispanic White ethnicity. More inconsistent and modest gender differences are apparent across cells, with females usually performing slightly better.

Age Effects

The Pearson correlation coefficients between age as a continuous variable and performance on the NIHTB-CB and Gold Standard composite measures are quite comparable and in accord with prior hypotheses: Fluid Cognition composites showed large, negative age effects (r (264) = − .68 and − .55 for NIHTB-CB and Gold Standard, p <.0001), whereas much smaller *positive* age effects were seen for the Crystallized Cognition composites (r = .18 and .14, p <.02), and the Total Cognition composites showed modest, negative effects (r = − .26 and − .22, p <.001). Steiger's Z tests confirm that the correlation between age and TB Fluid is indeed different from that between age and TB Crystallized (r = − 0.68 *vs.* r = +0.18, p <.001), and similar results were obtained comparing correlations between age and those respective GS composites (r = − 0.55 *vs.* r = +0.14, p <.001). Figures 1 and 2 show almost identical (overlapping) age trajectories for the NIHTB-CB and Gold Standard Fluid and Crystallized composites.

Other Demographic Differences

There were significant effects of education on NIHTB-CB Crystallized, Fluid, and Total Cognition Composite scores, after controlling for age and gender in the general linear model (GLM) analyses (see Table 5). As also shown in Table 5, after controlling for age and education, females scored somewhat higher than males on the NIHTB-CB and Gold Standard Fluid Composite scores, as well as NIHTB-CB Total Cognition Composite scores.

Race/ethnicity effects for self-described Caucasian (non-Hispanic White), African American (non-Hispanic Black) and Hispanic categories were examined separately from the GLM analyses, excluding the small number of multi-racial participants, and with age, education and gender being covaried in the analyses. Significant race/ethnicity effects were found for

the NIHTB-CB Crystallized Cognition Composite (F (2,252) = 21.58, p <.001), Fluid Cognition Composite (F (2,252) = 12.39, p < .001), and Total Cognition Composite scores (F (2,252) = 24.13, p < .001). In all cases, non-Hispanic Whites had higher scores than non-Hispanic Blacks and Hispanics. Very similar results were found for the Gold Standard composites, with respect to independent effects of education, gender and race/ethnicity.

There also were statistically significant effects of family income, with positive effect sizes (Cohen's d's) between the highest and lowest income groups ranging from 0.62 to 0.98 on the age-adjusted NIHTB-CB Crystallized Cognition Composite (F (3,245) = 9.32; p < .001), Fluid Cognition Composite (F (3,245) = 7.54; p < .001), and Total Cognition Composite scores (F (3,245) = 11.75; p < .001).

Relations with Prior School Difficulties and Current Health Status and Employment

Poorer performances on all three NIHTB-CB age-corrected (by covariance) cognitive composites were significantly related to self-reports of prior learning difficulties in school: repeating a grade (Crystallized p < .0001, Fluid p = .002, Total p < .0001), failing a grade (Crystallized p <0.001, Fluid p <.0001, Total p < .0001), special classes/tutoring (Crystallized p = 0.010, Fluid p = .049, Total p = .006), and overall school performance (ordered differences for above average > average > below average, p < .0001 for all TB composites). Again, similar results were found for the Gold Standard composites.

Reported better overall current health status also was significantly and positively related to the age-adjusted (by covariance) NIHTB-CB Crystallized Cognition Composite (F (2,261) = 8.85, p <.001), Fluid Cognition Composite (F (2,261) = 11.23, p <.001), and Total Cognition Composite (F (2,261) = 13.43, p <.001). In each case, the participants described as having "excellent" or "very good" health performed better than those described as having less than very good (poor to good) health, with mostly medium effect sizes (range, .26 to .84, median = .56). Self-reports of more specific health problems and health-related disability (diabetes, hypertension, lung/breathing problems) also were associated with worse, age-adjusted (by covariance) NIHTB-CB performances, again with small to medium effect sizes. Participants with diabetes evidenced worse Crystallized Cognition composites (F (1,262) = 5.56; p = .019; ES = − .47), Fluid Cognition composites (F (1,262) = 2.94; p = .088; ES =− .26), and Total Cognition composites (F (1,262) = 6.47; p = .012; ES = − .50). Those with reported hypertension had worse Fluid Cognition composites (F (1,262) = 5.48; p = .020; ES = − .25) and Total Cognition composites (F (1,262) = 2.88; p = .091; ES = − .24), and those with lung/breathing problems also had worse Crystallized Cognition composites (F (1,261) = 5.74; p = .017; ES = − .42), Fluid Cognition composites (F (1,262) = 3.12; p = .078; ES = − .23), and Total Cognition composites (F (1,262 =5.74; p = .017; ES = − .41). Finally, participants who endorsed having a health-related disability performed worse on Crystallized (F (1,263) = 3.65; p = .057; ES = − .32), Fluid (F (1,263 = 2.61; p = .108; ES = − .20), and Total (F (1,263) = 4.52; p = .034; ES = − .35) Cognition composites.

Employment status (employed or retired *vs*. out of work) was positively associated (small to medium effect sizes) with age-adjusted (by covariance) scores on the NIHTB-CB Crystallized Cognition Composite (F (1,239) = 4.29; p = .039; ES = .35), Fluid Cognition

Composite (F (1,239) = 3.65; p = .057; ES = .24), and Total Cognition Composite (F (1,239) = 5.14; p = .024; ES = .38).

DISCUSSION

The current study describes psychometric characteristics of the NIHTB Cognition Battery composite scores in a sample of adults, ages 20 to 85 years, who participated in the initial validation study. Similar psychometric data have been reported for the childhood portion of the validation sample (ages, 3 to 15 years; N = 208) (Akshoomoff, et al., 2013) and, therefore, were not included in this study. However, similarities and differences between the childhood and current adult sample findings will be noted below because they illustrate one of the major advantages of the NIHTB Cognition Battery (NIHTB-CB): namely, that it was designed to assess uniquely, with the same instruments, cognitive abilities across the human lifespan (ages 3 to 85 years).

As mentioned above, other, more statistically based (e.g., factor analytic) methods of deriving summary scores for cognitive test batteries exist. In fact, Mungas et al. (2013; and "submitted" for this series) have reported results of confirmatory factor analyses for the NIHTB-CB and related gold standard measures, in both child and adult samples. Such approaches are particularly helpful in identifying and analyzing the specific constructs that are measured by the battery at different ages. Three distinct NIHTB-CB dimensions (Vocabulary, Reading, and Fluid Abilities) were identified with the youngest children (ages 3–6 years), whereas five or six were seen with older children and adults (Vocabulary, Reading, Episodic Memory, Working Memory, and Executive Function/ Processing Speed). In addition, however, second order factors consistent with the current crystallized and fluid categories parsimoniously accounted for the correlations among first order factors. When more fine-grained analyses of NIHTB-CB domains are needed, the factors identified by Mungas and colleagues may be preferred. On the other hand, when a smaller number of cognitive outcome variables is desired, and less precision is required, the current Crystallized and Fluid (or even Total) composites have advantages of greater parsimony, theoretical relevance and validity across the entire lifespan, as well as adequate internal consistencies.

As in the child study, we found high test–retest reliabilities of the NIHTB-CB composite scores with adults (r = .88–.92). Together, these results support use of the NIHTB-CB composites for tracking cognition across the lifespan in future longitudinal research. Also similar to the results reported for children, no significant practice effects were noted for adults with repeated administrations of the NIHTB-CB Crystallized composite; however, the Fluid and Total Cognition composites did show practice effects. Given that the NIHTB-CB is a "same-version" repeated battery, the practice effects observed are not unexpected for such same-version tests that involve adaptation to novel requirements as opposed to those that mainly tap past accumulated semantic knowledge (Fluid *vs*. Crystallized abilities). In fact, the NIHTB-CB practice effects were similar to those obtained for the widely used gold standard measures. However, this is a departure from equivalent alternate forms that are available for some neuropsychological tests (e.g., the Repeatable Battery for Assessment of Neuropsychological Status; RBANS; Randolph, 1998), which tend to minimize practice

effects even on fluid cognition measures across time. As such, when interpreting NIHTB-CB performances on second or subsequent assessments, appropriate test–retest adjustments would need to be made as are applied with other same-version neuropsychological batteries in longitudinal contexts (e.g., reliable change indices or regression-based norms that adjust for test–retest reliability, practice effects, demographics and other factors that may affect test–retest differences; Cysique et al., 2011).

In our previous report on the childhood sample (Akshoomoff et al., 2013), NIHTB-CB Crystallized and Fluid Cognition composites showed strong, linear improvements from ages 3 to 15 years. Indeed, the child sample's mean Crystallized and Fluid positive age trajectories were virtually overlapping. During adulthood, however, age trajectories for these composites were quite different (see Table 5 and Figures 1 and 2). Overall, the NIHTB-CB and Gold Standard *Crystallized* Cognition composites show no significant, independent age effect in normal adults (Table 5). On the other hand, as expected, both NIHTB-CB and Gold standard *Fluid* Cognition composites show strong, linear, age-related decline from young to old adulthood (Figure 2). This pattern is consistent with what has been observed with other cognitive tests and composites that reflect fluid constructs of processing speed, attention, working memory, episodic memory, and executive function (Heaton et al., 2003, 2004, 2005).

Also as expected, the NIHTB-CB and Gold Standard Crystallized Cognition composites showed by far the strongest relationships with education (Table 5), as they reflect prior learning of language and reading skills in school and elsewhere. Fluid Cognition composites also show significant, albeit smaller, independent education effects, possibly because there is a tendency for people who are more generally cognitively able to go further in school, and to enter jobs/professions that require (and practice) cognitive abilities of a "fluid" nature (Matarazzo, 1972).

Consistent with prior findings with children, there were no independent gender effects on the NIHTB-CB Crystallized Cognition composite (Table 5). Modest effects, favoring women, were seen for adults on the NIHTB-CB Fluid and Total Cognition composites, however. This type of small but statistically significant gender difference has been noted on other tests, particularly those that tap processing speed and episodic memory (e.g., on Wechsler Intelligence and Memory Scales; Heaton et al., 2003). Also consistent with NIHTB-CB findings with children, as well as with results from adult samples on many published cognitive tests (e.g., Norman et al., 2011), substantial ethnicity effects (favoring Caucasians) were found here on all NIHTB-CB composites. These ethnicity effects undoubtedly have multiple, complex causes that relate to socioeconomic backgrounds, quality of education (not captured in "years completed"), and other potential disadvantages associated with growing up as a member of a minority group in a developed country (e.g., Byrd et al., 2006).

An important goal of future research is to clarify specific *causes* of all demographic differences in cognitive test performance (those related to age, education, and gender, as well as ethnicity). Regardless of specific causes, however, it is clear that such effects, in aggregate, are substantial on virtually all cognitive tests, and should be considered in

establishing normative standards for determining whether there has been any *change* in a person's cognitive functioning (e.g., due to disease or illness involving the brain).

U.S. national norming of the NIHTB Cognition Battery is currently under way for both English and Spanish speakers. The planned norms will permit the user to choose between demographic corrections: (1) for age alone, when one wants to compare an individual's or group's results with normal expectations for the general U.S. population at any given age group; or (2) for all demographic effects that have been found to significantly affect cognition in normal individuals, including those that may relate to age, level of education, gender, or ethnicity (non-Hispanic Caucasian, non-Hispanic African American, Hispanic). In addition, a Spanish version of the battery has been developed and will also undergo norming in anticipation of growing numbers of Hispanics and Latinos participating in research studies.

Co-administration of the NIHTB Cognition Battery with well-established (Gold Standard) tests of the same crystallized and fluid abilities enabled us to assess convergent and discriminant (construct) validity of the NIHTB-CB composites. When this was done previously with children ages 3 to 15 years, we found strong evidence of convergent validity (very high correlations between NIHTB-CB and Gold Standard Crystallized Cognition composites, and between NIHTB-CB and Gold Standard Fluid Cognition composites); however, with children, discriminant validity was difficult to establish because the powerful developmental effect resulted in high correlations among both types of abilities in this age range. On the other hand, with adults we were able to demonstrate both convergent and discriminant validity of the NIHTB-CB composites: each specific NIHTB-CB composite score (Crystallized and Fluid) correlated very strongly with its Gold Standard counterpart, and much less so with the other specific Gold Standard composite. These findings must be considered preliminary because many potential Gold Standard tests were not included in the current study (e.g., the latest, full versions of the Wechsler Intelligence and Memory Scales), but the current results do support the construct validity of the NIHTB-CB composites.

Even more preliminary is the currently available evidence of external/predictive validity of the NIHTB-CB summary measures, because that evidence lacks desired specificity and is totally based upon self-report. Consistent with previously reported associations between children's NIHTB-CB composites and their mothers' ratings of contemporaneous academic functioning (Akshoomoff et al., 2013), the adults in the current sample who reported worse overall prior school performance, and/or having repeated or failed a grade, or having required special class placement or tutoring, performed significantly worse on the NIHTB-CB (especially the Crystallized and Total Cognition composites). Those who reported less than "very good" current health status also performed worse on the NIHTB-CB (especially the Fluid and Total Cognition composites). Finally, self-reports of having certain risks for compromised brain function (hypertension, diabetes, and lung/breathing problems) also were associated with significantly worse scores on the NIHTB-CB composites. All these associations between NIHTB-CB composite scores and self-reported health status and everyday functioning (including prior academic functioning, current disability and employment status) suggest that the NIHTB-CB may be useful in both epidemiologic research and studies of clinical conditions in which cognitive outcomes are considered

important. Nevertheless, further research is needed to establish the NIHTB Cognition Battery's sensitivity to verified health conditions, including longitudinal research that determines whether the NIHTB-CB measures are sensitive to clinically significant changes in brain function over time.

The NIHTB Cognition Battery's computer-based presentation and automatic scoring and norming arguably increase both efficiency and accuracy of testing, and reduce the need for extensive examiner training and expertise. However, the examiner still has a critical role in ensuring that the test taker understands the standard instructions and consistently puts forth adequate effort. This role is likely to be particularly important when examining individuals who may have cognitive or emotional conditions that could compromise validity of the assessment.

In conclusion, especially when considered together with previously reported data on a childhood validation sample (Akshoomoff et al., 2013), the current findings with adults suggest that the NIHTB-CB Crystallized, Fluid and Total Cognition composite scores are highly reliable, have good construct validity, and are likely to be useful in tracking clinically and epidemiologically relevant cognitive outcomes across the lifespan.

Acknowledgments

We thank Abigail Sivan and Edmond Bedjeti (Northwestern University) for their valuable assistance in the validation phase of testing. We also thank the following individuals for their helpful consultation during the development of the NIH Toolbox Cognition Battery: Jean Berko Gleason (Boston University), Rachel Byrne (Kessler Foundation), Gordon Chelune (University of Utah), Nancy Chiaravalotti (Kessler Foundation), Dean Delis (University of California, San Diego), Adele Diamond (University of British Columbia), Roberta Golinkoff (University of Delaware), Kathy Hirsh-Pasek (Temple University), Marilyn Jager Adams (Brown University), Joel Kramer (University of California, San Francisco), Joanie Machamer (University of Washington), Amanda O'Brien (Kessler Foundation), Timothy Salthouse (University of Virginia), Jerry Sweet (University of Chicago), Keith O. Yeates (Ohio State University), and Frank Zelkoe (Northwestern University). This study is funded in whole or in part with Federal funds from the Blueprint for Neuroscience Research, National Institutes of Health, under Contract No. HHS-N-260-2006-00007-C. We have no conflicts of interest to report. Dr. Heaton is funded by NIH grants # P30MH062512, P50DA026306, P01DA012065, R01MH060720, R01MH073433, R01MH058076, R01MH078748, R01MH078737, U01MH083506, R01MH083552, R01MH081861. Dr. Akshoomoff reports no disclosures. Dr. Tulsky is funded by NIH contracts H133B090024, H133N060022, H133G070138, B6237R, cooperative agreement U01AR057929, and grant, R01HD054659. He has received consultant fees from the Institute for Rehabilitation and Research, Frazier Rehabilitation Institute/Jewish Hospital, Craig Hospital, and Casa Colina Centers for Rehabilitation. Dr. Mungas is funded by research grants from the National Institute on Aging and a grant from the California Department of Public Health California Alzheimer's Disease Centers program. Dr. Weintraub is funded by NIH grants # R01DC008552, P30AG013854, and the Ken and Ruth Davee Foundation and conducts clinical neuropsychological evaluations (35% effort) for which her academic-based practice clinic bills. She serves on the editorial board of Dementia & Neuropsychologia and advisory boards of the Turkish Journal of Neurology and Alzheimer's and Dementia. Dr. Dikmen receives research grant funding from NIH R01 NS058302 and R01HD061400, NIDRR H133A080035, NIDRR H133G090022, and NIDRR, H133A980023, and DoD W81XWH-0802-0159. Ms. Beaumont served as a consultant for NorthShore University HealthSystem, FACIT.org, and Georgia Gastroenterology Group PC. She received funding for travel as an invited speaker at the North American Neuroendocrine Tumor Symposium. Ms. Casaletto is supported by NIH grants F31-DA035708 and T32-DA31098. Dr. Conway reports no disclosures. Dr. Slotkin reports no disclosures. Dr. Gershon has received personal compensation for activities as a speaker and consultant with Sylvan Learning, Rockman, and the American Board of Podiatric Surgery. He has several grants awarded by NIH: N01-AG-6-0007, 1U5AR057943-01, HHSN260200600007, 1U01DK082342-01, AG-260-06-01, HD05469, NINDS: U01 NS 056 975 02, NHLBI K23: K23HL085766 NIA; 1RC2AG036498-01; NIDRR: H133B090024, OppNet: N01-AG-6-0007. Disclaimer: The views and opinions expressed in this report are those of the authors and should not be construed to represent the views of NIH or any of the sponsoring organizations, agencies, or the U.S. government.

REFERENCES

Akshoomoff, N.; Beaumont, JL.; Bauer, PJ.; Dikmen, S.; Gershon, R.; Mungas, D.; Heaton, RK. NIH Toolbox Cognitive Function Battery (CFB): Composite scores of crystallized, fluid, and overall cognition. In: Zelazo, PD.; Bauer, PJ., editors. National Institutes of Health Toolbox-Cognitive Function Battery (NIH Toolbox CFB): Validation for Children between 3 and 15 years. Society for Research in Child Development Monographs. Vol. 78. 2013. p. 199-132.

Benedict, R. Brief Visuospatial Memory Test-revised. Odessa, FL: Psychological Assessment Resources, Inc.; 1997.

Byrd DA, Miller SW, Reilly J, Weber S, Wall TL, Heaton RK. Early environmental factors, ethnicity, and adult cognitive test performance. The Clinical Neuropsychologist. 2006; 20:243–260. [PubMed: 16690545]

Cattell, RB. Abilities: Their structure, growth, and action. Cambridge: Cambridge University Press; 1971.

Cohen J. A power primer. Psychological Bulletin. 1992; 112:155–159. [PubMed: 19565683]

Cysique L, Franklin D, Abramson I, Ellis R, Letendre S, Collier A, Heaton RK. Normative data and validation of a regression based summary score for assessing meaningful neuropsychological change. Journal of Clinical and Experimental Neuropsychology. 2011; 33(5):505–522. [PubMed: 21391011]

Delis, DC.; Kramer, JH.; Kaplan, E. The Delis-Kaplan Executive Function System. San Antonio, TX: The Psychological Corporation; 2001.

Dunn, LM.; Dunn, LM. Peabody Picture Vocabulary Test-Fourth Edition (PPVT-4). Circle Pines, MN: American Guidance Services; 2007.

Gronwall DM. Paced auditory serial-addition task: A measure of recovery from concussion. Perceptual and Motor Skills. 1977; 44:367–373. [PubMed: 866038]

Heaton, RK.; Chelune, GJ.; Talley, JL.; Kay, GG.; Curtiss, G. Wisconsin card sorting test manual. Odessa FL: Psychological Assessment Resources, Inc.; 2004.

Heaton, RK.; Miller, SW.; Taylor, JT.; Grant, I. Revised comprehensive norms for an expanded Halstead-Reitan Battery: Demographically adjusted neuropsychological norms for African American and caucasian adults. Lutz, FL: Psychological Assessment Resources, Inc.; 2004.

Heaton, RK.; Ryan, L.; Grant, I. Demographic influences and use of demographically corrected norms in neuropsychological assessment. In: Grant, I.; Adams, KM., editors. Neuropsychological assessment of neuropsychiatric and neuromedical disorders. New York: Oxford University Press; 2009. p. 127-155.

Heaton, RK.; Taylor, MJ.; Manly, J. Demographic effects and use of demographically corrected norms with the WAIS-III and WMS III. In: Tulsky, DS.; Saklofske, DH.; Chelune, GJ.; Heaton, RK.; Ivnik, RJ.; Bornstein, R.; Prifitera, A.; Ledbetter, MF., editors. Clinical interpretation of the WAIS-III and WMS-III. San Diego, CA: Academic Press; 2003. p. 181-210.

Horn JL. Organization of abilities and the development of intelligence. Psychological Review. 1968; 75:242–259. [PubMed: 4875815]

Horn, JL. Organization of data on life-span development of human abilities. In: Goulet, LR.; Baltes, PB., editors. Life-span developmental psychology: Research and theory. San Diego: Academic Press; 1970. p. 423-466.

Horn JL, Cattell RB. Age differences in fluid and crystallized intelligence. Acta Psychologica. 1967; 26:107–129. [PubMed: 6037305]

Li S-C, Lindenberger U, Hommel B, Aschersleben G, Prinz W, Baltes P. Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. Psychological Science. 2004; 15:155–163. [PubMed: 15016286]

Matarazzo, JD. Wechsler's measurement and appraisal of adult intelligence. 5th ed.. Baltimore, MD: The Williams and Wilkins Company; 1972.

Mungas D, Heaton RK, Tulsky D, Zelazo PD, Slotkin J, Blitz D, Gershon R. Factor structure, convergent validity, and discriminant validity of the NIH Toolbox Cognitive Health Battery (NIHTB-CHB) in Adults. Journal of the International Neuropsychological Society. (submitted).

Mungas, D.; Widaman, K.; Zelazo, PD.; Tulsky, D.; Heaton, R.; Slotkin, J.; Gershon, R. NIH Toolbox Cognitive Function Battery (CFB): Factor structure for 3- to 15- year-olds. In: Zelazo, PD.; Bauer, PJ., editors. National Institutes of Health Toolbox-Cognitive Function Battery (NIH Toolbox CFB):validation for children between 3 and 15 years. Ann Arbor, MI: Society for Research in Child Development Monographs. Vol. 78. 2013. p. 103-118.

Norman MA, Moore DJ, Taylor M, Franklin D, Cysique L, Ake C, Heaton RK. Demographically corrected norms for African Americans and Caucasians on the Hopkins Verbal Learning Test-Revised, Brief Visuospatial Memory Test- Revised, Stroop Color and Word Test, and Wisconsin Card Sorting Test 64-Card Version. Journal of Clinical and Experimental Neuropsychology. 2011; 33(7):793–804. [PubMed: 21547817]

Randolph, C., editor. Repeatable Battery for the Assessment of Neuropsychological Status Manual. San Antonio, TX: The Psychological Corporation; 1998.

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

Sattler, JM. Assessment of children: Cognitive applications. San Diego, CA: Jerome M. Sattler, Publisher, Inc.; 2001.

The Psychological Corporation. WAIS-III WMS-III Technical Manual. San Antonio, TX: The Psychological Corporation; 1997.

Wechsler, D. Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV). San Antonio, TX: The Psychological Corporation; 2008.

Weintraub S, Dikmen SS, Heaton RK, Tulsky DS, Zelazo PD, Bauer PJ, Gershon R. Cognition assessment using the NIH Toolbox. Neurology. 2013; 80:S54–S64. [PubMed: 23479546]

Wilkinson, GS.; Robertson, GJ. Wide Range Achievement Test 4 professional manual. Lutz, FL: Psychological Assessment Resources; 2006.

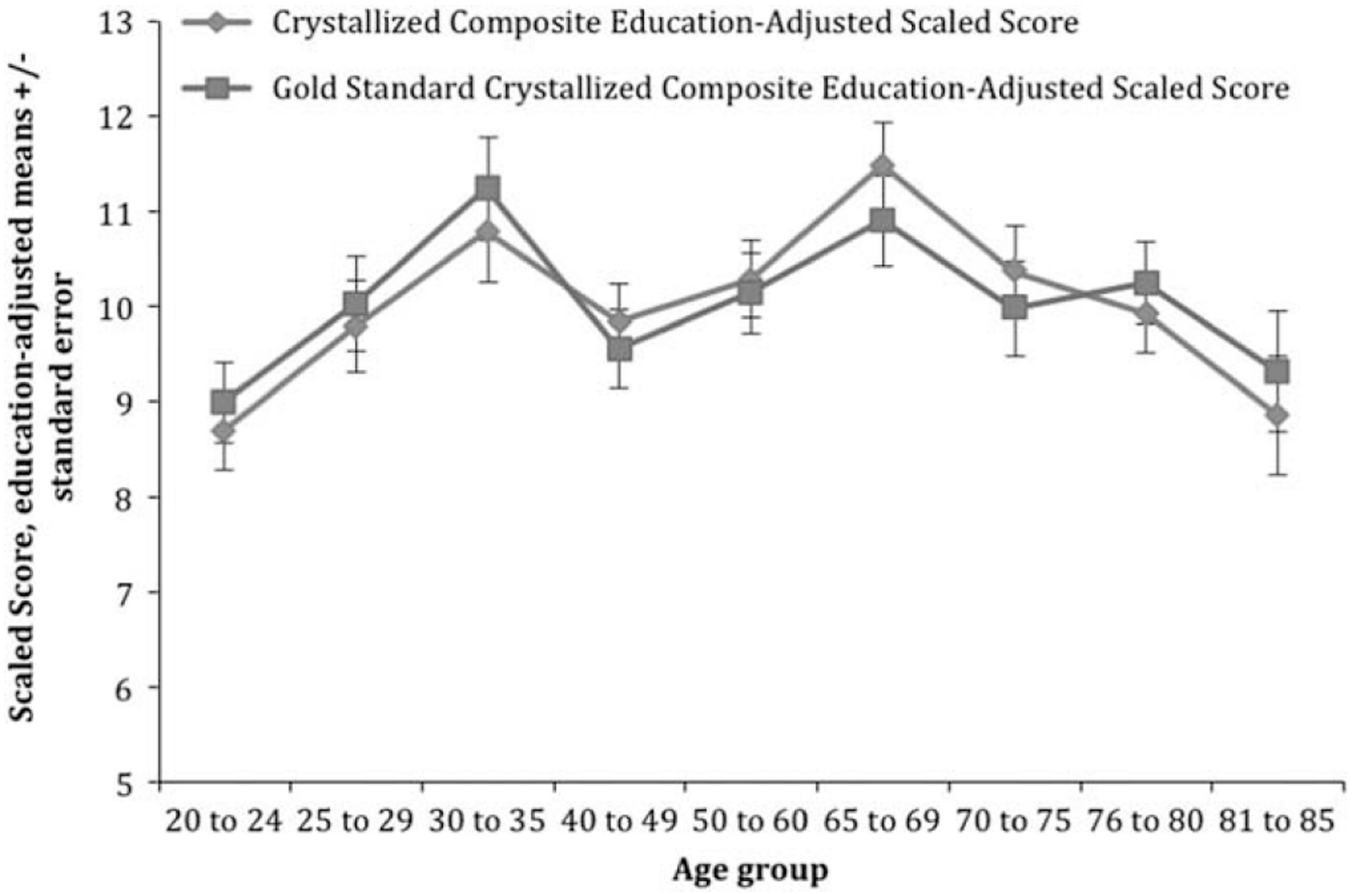

Fig. 1.

Performance on the Toolbox Crystallized Cognition Composite and the Gold Standard Crystallized Cognition Composite across age groups, adjusted for education. Error bars represent±1 SE.

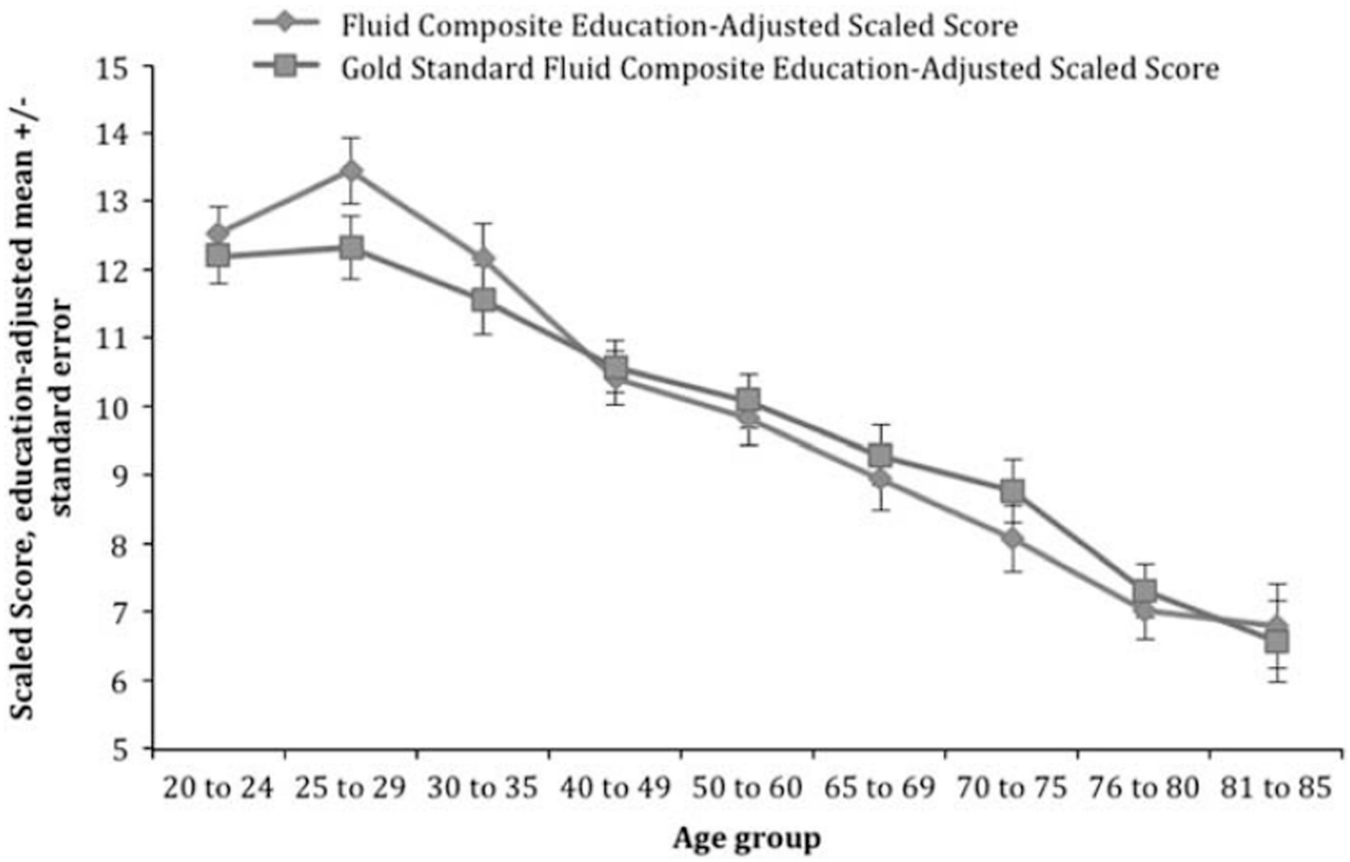

Fig. 2.

Performance on the Toolbox Fluid Cognition Composite and the Gold Standard Fluid Cognition Composite across age groups, adjusted for education. Error bars represent ±1 SE.

Table 1

Intraclass (*ICC*) and Pearson's (*r*) correlations per NIH Toolbox and Gold Standard composite scores

	Crystallized	Fluid	Total
Toolbox			
Pearson's r (p-value)	0.92	0.86	0.90
	(*p* <.001)	(*p*<.001)	(*p* <.001)
ICC (95% CI)	0.92	0.79	0.86
	(0.88, 0.95)	(0.70, 0.86)	(0.80, 0.91)
Gold Standard			
Pearson's r (p-value)	0.93	0.95	0.95
	(*p* <.001)	(*p*<.001)	(*p* <.001)
ICC (95% CI)	0.92	0.88	0.90
	(0.88, 0.95)	(0.82, 0.92)	(0.85, 0.93)

Table 2

NIH Toolbox test-retest practice effect sizes (*N* = 89; mean interval 15.5 days, *SD* = 4.8)

Test	Time 1 Mean (*SD*)	Time 2 Mean (*SD*)	*t*	*p*-value	Cohen's d
TB Crystallized	9.9 (3.0)	10.0 (3.0)	1.23	.222	0.05
GS Crystallized	9.8 (3.0)	10.1 (3.0)	2.36	.021	0.10
TB Fluid	10.2 (3.0)	11.4 (3.1)	7.31	<.001	0.42
GS Fluid	10.0 (2.8)	11.2 (3.1)	11.1	<.001	0.40
TB Total	10.0 (2.9)	10.8 (3.1)	6.03	<.001	0.29
GS Total	9.9 (2.9)	10.7 (3.2)	8.09	<.001	0.30

Note. TB = NIH Toolbox; GS = Gold Standard.

Table 3

Correlations among uncorrected Composite Scores for NIH Toolbox (TB) and Gold Standard (GS) batteries: Evidence for convergent and discriminant validity

	TBC	GSC	TBF	GSF	TBT	GST
TBC	—	0.90	0.17	0.39	0.80	0.78
GSC	0.90	—	0.19	0.39	0.75	0.85
TBF	0.17	0.19	—	0.78	0.71	0.56
GSF	0.39	0.39	0.78	—	0.74	0.81
TBT	0.80	0.75	0.71	0.74	—	0.89

Note. C = Crystallized; F = Fluid; T = Total.

J Int Neuropsychol Soc. Author manuscript; available in PMC 2015 July 01.

Table 4

NIH Toolbox Total Cognition Composite scores by age, education, gender, and ethnicity: mean, (*SD*), and range

		Gender		Race/ethnicity		
Age group	**Education**	**Male**	**Female**	**White**	**Black**	**Hispanic/other**
20–60 Yrs	<High school	*n* = 22	*n* = 26	*n* = 21	*n* = 15	*n* = 12
		9.5 (2.2)	9.9 (3.4)	11.7 (2.6)	7.7 (2.0)	8.9 (2.3)
		4.9–13.9	3.9–15.7	6.6–15.7	4.1–9.9	3.9–11.6
	High school gradate	*n* = 29	*n* = 31	*n* = 26	*n* = 19	*n* = 15
		10.3 (2.7)	9.9 (2.3)	10.9 (2.4)	9.5 (2.1)	9.5 (2.8)
		5.9–16.3	4.5–13.7	7.4–14.8	5.9–13.1	4.5–16.3
	≥College	*n* = 24	*n* = 27	*n* = 24	*n* = 15	*n* = 12
		11.3 (1.9)	12.0 (2.5)	11.9 (2.2)	11.1 (1.6)	11.8 (2.7)
		8.1–15.9	8.2–18.5	8.2–18.5	8.1–13.6	8.3–15.9
65–85 Yrs	<High school	*n* = 9	*n* = 11	*n* = 9	*n* = 10	*n* = 1
		5.5 (1.9)	6.5 (3.0)	7.6 (2.1)	5.0 (2.3)	2.5
		2.5–9.3	1.5–10.7	5.2–10.7	1.5–9.3	
	High school gradate	*n* = 12	*n* = 27	*n* = 26	*n* = 11	*n* = 2
		8.5 (2.5)	9.1 (2.8)	9.4 (2.5)	7.2 (2.3)	11.6 (3.9)
		4.3–11.9	5.1–16.6	5.8–16.6	4.3–11.0	8.8–14.4
	≥College	*n* = 23	*n* = 26	*n* = 41	*n* = 5	*n* = 3
		10.2 (2.9)	11.5 (2.9)	11.4 (2.7)	8.0 (3.4)	8.0 (1.8)
		5.0–16.1	6.6–17.5	5.0–17.5	5.3–13.9	6.6–10.0
TOTAL		*n* = 119	*n* = 148	*n* = 147	*n* = 75	*n* = 45
		9.8 (2.8)	10.2 (3.1)	10.9 (2.7)	8.4 (2.8)	9.8 (3.0)
		2.5–16.3	1.5–18.5	5.0–18.5	1.5–13.9	2.5–16.3

J Int Neuropsychol Soc. Author manuscript; available in PMC 2015 July 01.

Table 5

General linear models reflecting the independent effects of age, education and gender on demographically uncorrected NIH Toolbox and Gold Standard Composite scores

	Model	Regression coefficient	Standard error	*t* value	*p*-Value
Crystallized Composite Toolbox					
Adjusted R^2	0.274				
F	32.99				
Gender = female		0.377	0.319	1.18	.238
Age, years		0.011	0.008	1.47	.142
Education, years		0.512	0.055	9.29	<.001
Crystallized Composite Gold Standard					
Adjusted R^2	0.249				
F	28.84				
Gender = female		−0.260	0.325	−0.80	.424
Age, years		0.008	0.008	1.07	.285
Education, years		0.495	0.056	8.83	<.001
Fluid Composite Toolbox					
Adjusted R^2	0.501				
F	87.75				
Gender = female		0.644	0.265	2.43	.016
Age, years		−0.104	0.006	−16.20	<.001
Education, years		0.180	0.046	3.95	<.001
Fluid Composite Gold Standard					
Adjusted R^2	0.380				
F	53.39				
Gender = female		0.652	0.295	2.21	.028
Age, years		−0.088	0.007	−12.31	<.001
Education, years		0.264	0.051	5.19	<.001
Total Cognition Composite Toolbox					
Adjusted R^2	0.269				
F	32.07				

J Int Neuropsychol Soc. Author manuscript; available in PMC 2015 July 01.

	Model	Regression coefficient	Standard error	*t* value	*p*-Value
Gender = female		0.680	0.320	2.12	.035
Age, years		−0.052	0.008	−6.68	<.001
Education, years		0.457	0.055	8.26	<.001
Total Cognition Composite Gold Standard					
Adjusted R^2	0.235				
F	26.69				
Gender = female		0.255	0.328	0.78	.438
Age, years		−0.042	0.008	−5.36	<.001
Education, years		0.456	0.057	8.06	<.001