NIH Toolbox Cognitive Function Battery (CFB): Composite Scores of Crystallized, Fluid, and Overall Cognition

Natacha Akshoomoff[1], **Jennifer L. Beaumont**[2], **Patricia J. Bauer**[3], **Sureyya Dikmen**[4], **Richard Gershon**[2], **Dan Mungas**[5], **Jerry Slotkin**[2], **David Tulsky**[6], **Sandra Weintraub**[7], **Philip Zelazzo**[8], and **Robert K. Heaton**[1]

[1]Department of Psychiatry, University of California, San Diego, CA

[2]Department of Medical Social Sciences, Northwestern University, Chicago, IL

[3]Department of Psychology, Emory University, Atlanta, Georgia

[4]Department of Rehabilitation Medicine, University of Washington, Seattle, Washington

[5]Department of Neurology, University of California, Davis, CA

[6]Department of Physical Medicine and Rehabilitation, University of Michigan, Ann Arbor, Michigan

[7]Cognitive Neurology and Alzheimer's Disease Center and Departments of Psychiatry and Neurology, Northwestern Feinberg School of Medicine, Chicago, Illinois

[8]Institute of Child Development, University of Minnesota, Minneapolis, Minnesota

Abstract

The NIH Toolbox Cognitive Function Battery (CFB) includes 7 tests covering 8 cognitive abilities considered to be important in adaptive functioning across the lifespan (from early childhood to late adulthood). Here we present data on psychometric characteristics in children ($N = 208$; ages 3–15 years) of a total summary score and composite scores reflecting two major types of cognitive abilities: "crystallized" (more dependent upon past learning experiences) and "fluid" (capacity for new learning and information processing in novel situations). Both types of cognition are considered important in everyday functioning, but are thought to be differently affected by brain health status throughout life, from early childhood through older adulthood. All three Toolbox composite scores showed excellent test-retest reliability, robust developmental effects across the childhood age range considered here, and strong correlations with established, "gold standard" measures of similar abilities. Additional preliminary evidence of validity includes significant associations between all three Toolbox composite scores and maternal reports of children's health status and school performance.

The NIH Toolbox Cognitive Function Battery (CFB), together with test modules for motor, sensory, and emotional functioning, comprise the "NIH Toolbox for Neurological and Behavioral Function." The development of the NIH Toolbox was commissioned by multiple

Correspondence may be addressed to Robert K. Heaton, Department of Psychiatry, 9500 Gilman Drive, University of California, San Diego, La Jolla, CA, 92093-0603, USA; rheaton@ucsd.edu.

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). Chapter 1 in this monograph describes the rationale for the NIH Toolbox, a NIH Blueprint initiative. That chapter also describes the methods for identifying target subdomains and appropriate test instruments.

The NIH Toolbox CFB is currently comprised of 7 test instruments that measure 8 abilities within 6 major cognitive domains. The test instruments are described in detail in Chapters 2 through 6 of this monograph: Chapter 2—the Dimensional Change Card Sort (DCCS) Test (Executive Function-Cognitive Flexibility) and the Flanker Inhibitory Control and Attention Test (Executive Function-Inhibitory Control and Sustained Attention); Chapter 3—the Picture Sequence Memory Test (Episodic Memory-Visual); Chapter 4—the Picture Vocabulary Test (Language-Vocabulary Comprehension), and the Oral Reading Recognition Test (Language-Reading Decoding); Chapter 5—the List Sorting Working Memory Test (Working Memory); and Chapter 6—the Pattern Comparison Processing Speed Test.

Many researchers will want to consider measures of these various cognitive functions separately, but others are expected 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., this volume) but these may 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 adult).

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. 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 is dynamic, developing and transforming 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 presumed to be especially influenced by biological processes and less dependent on past exposure (learning experiences). These abilities improve rapidly during childhood, typically reaching their peak in early adulthood, and then decline as adults get older. Crystallized abilities, in contrast, are presumed to be more dependent on experience, and less by biological influences. They represent accumulated store of verbal knowledge and skills, and thus are more heavily influenced by education and cultural exposure, particularly during childhood. These abilities show marked developmental change during childhood, but they typically continue to improve slightly into middle adulthood and then remain relatively stable.

Age-related improvements in fluid abilities in early development are thought to support acquisition of the knowledge needed for crystallized abilities, thus accounting for stronger

correlations between fluid and crystallized abilities early in life, compared with those found in later years (Cattell, 1971; Horn 1968). Once developed, crystallized abilities tend to be fairly stable throughout adulthood and much less susceptible to the effects of aging and health status during aging than is the case with fluid abilities. In contrast, fluid abilities tend to be more sensitive to neurobiological integrity, including changes in brain functioning with aging and in a variety of neurological disorders that alter brain structure and function.

Here we present data from the children's validation sample for the NIH Toolbox CFB that is based on three candidate summary scores: Toolbox Crystallized Cognition Composite, Toolbox Fluid Cognition Composite, and Toolbox Cognitive Function Composite (a combination of both crystallized and fluid scores). Results are based on 208 children, ages 3 to 15. We expected all three summary scores to increase fairly rapidly with age, in contrast to results obtained during adulthood (Weintraub et al., submitted). We also present psychometric information, such as 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). Although we predicted that the NIH Toolbox CFB summary scores would show good convergent validity with relevant gold standard measures, we expected that there would be less evidence of discriminant validity across fluid and crystallized abilities, particularly among younger children. This hypothesis was based upon the expectation that fluid and crystallized abilities develop rapidly and roughly in parallel during early childhood, whereas 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, 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 may also differ with respect to gender, family income, and race/ethnicity. Whereas level of formal education also is a significant predictor of cognitive test performance in adulthood (e.g., Heaton et al., 2003, 2004), in children age and education are almost totally confounded. However, for children, maternal level of education also has been shown to be a significant predictor of IQ and various aspects of neuropsychological test performance. The relation of each of these demographic variables with the composite measures of NIH Toolbox CFB performance were examined.

Finally, to further explore validity of the Toolbox composite measures, we examined associations between all cognitive summary scores and a few relatively gross measures of health and everyday functioning (maternal reports of health and school performance).

Method

Participants

The sample included 208 participants: 120 were ages 3 to 6 and 88 ages 8 to 15; 104 were females and 104 males; race/ethnicity composition of the sample was 92 Caucasian [non-Hispanic White], 65 African American, 42 Hispanic, and 9 multiracial (the 9 multiracial children were excluded from ethnicity comparisons due to the small sample size and greater heterogeneity). Maternal level of education was categorized as less than high school graduate (15%), high school graduate or some college (54%), and Bachelor's degree or higher (31%).

Additional demographic, health status and school functioning variables were based on the categorical information obtained from each participant's parent (typically, the mother). Family income was categorized into five levels (< $20,000[aa%], $20,000 to $39,999[bb%], $40,000 to $74,999[cc%], $75,000 to $99,999[dd%], and $100,000[ee%]). Child health status was categorized as Excellent (69%) or Poor to Very Good (31%). For school-age children (age 8–15) maternal ratings of academic performance were classified as Above Average (XX%) or Below Average to Average (YY%); also ZZ% of these children were reported to have required special (remedial) classes and/or tutoring in school.

Participants were recruited from 4 testing sites: 60 at Chicago's NorthShore University HealthSystems, 58 at Emory University in Atlanta, 17 at New Jersey's Kessler Institute for Rehabilitation, and 73 at the University of Minnesota. A subset of 66 participants (approximately 33%) completed a retest 7 to 21 days later to assess test-retest reliability and "practice effects."

NIH Toolbox CFB Measures

The entire battery of seven NIH Toolbox CFB tests was included in this study. This resulted in two measures of crystallized abilities (the Picture Vocabulary Test and the Oral Reading Recognition Test), as well as six measures of fluid abilities (the Dimensional Change Card Sort (DCCS) Test, the Flanker Inhibitory Control and Attention Test, the Picture Sequence Memory Test, the List Sorting Working Memory Test, and the Pattern Comparison Processing Speed Test). Descriptions of the individual NIH Toolbox CFB tests and the derived scores that reflect the multiple domains of cognitive functioning are provided in Chapters 2 to 6 of this monograph. Raw scores from the CFB 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 Toolbox Crystallized Cognition Composite, Toolbox Fluid Cognition Composite, and Toolbox Cognitive Function (i.e., Total) Composite scores. Three additional summary scores were computed in order to evaluate the use of a potential abbreviated version of the CFB with children ages 3 to 6 years: the Short Toolbox Crystallized Cognition Score (the Picture Vocabulary Test), the Short Toolbox Fluid Cognition Score (the Flanker Inhibitory Control and Attention Test [including measures of both Executive Function-Inhibitory Control and Sustained Attention] and the Picture Sequence Memory Test), and the Short Toolbox Total Composite that was a combination of the abbreviated crystallized and fluid scores.

"Gold Standard" Cognitive Measures

In Chapter 1, Tables 3 and 4 show the NIH Toolbox CFB instruments and analogous gold standard measures for each. Normalized standard scores (scaled scores) from two published and widely used measures, the Reading subtest from the Wide Range Achievement Test – 4th Edition (Wilkinson & Robertson, 2006) and the Peabody Picture Vocabulary Test – Fourth Edition (PPVT-IV; Dunn & Dunn, 2007), were combined for the Gold Standard Crystallized Composite score. Data available from gold standard fluid cognition measures varied with participants' age, because we are unaware of any previously published instruments that have been standardized for the full age range designated for the NIH Toolbox. In order to compare the NIH Toolbox CFB Fluid Composite with gold standard measures for children ages 3 to 6 years, scaled scores from the Wechsler Preschool and Primary Scale of Intelligence-Third Edition Block Design (Wechsler, 2002) and Sentence Repetition from the Developmental Neuropsychological Assessment, second edition (NEPSY-II; Korkman, Kirk, & Kemp, 2007) were combined for the Gold Standard Fluid Composite score. For children ages 8 to 15, the Gold Standard Fluid Composite score was derived from normalized scaled scores from the Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV) Letter-Number Sequencing subtest (Wechsler, 2003), an average of scores from the WISC-IV Coding and Symbol Search subtests (Wechsler, 2003), the Delis-Kaplan Executive Function System (Delis, Kramer & Kaplan, 2001) Color-Word Interference score, 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 the Paced Auditory Serial Addition Test (Gronwall, 1977; first channel only). In order to evaluate the Short Toolbox scores for the 3- to 6-year-olds, the Short Gold Standard Crystallized score was based on the PPVT-IV score. Also for 3- to 6-year-olds, Gold Standard Fluid scores (combination of WPPSI-III Block Design and NEPSY Sentence Repetition) were compared with Short Toolbox Fluid Cognition scores.

Analyses

Non-age-adjusted, normalized scaled scores were computed for each NIH Toolbox CFB measure. These were then averaged together to create normalized, composite scaled scores. Using data from the subset of participants who had repeated testing ($n = 66$, including $n = 38$ for ages 3–6 years and $n = 28$ for ages 8–15 years), Pearson correlations were computed to estimate test-retest reliability for the total retested group and for the age subgroups (3 to 6 years; 8 to 15 years) separately. The relation between each of the non-age-adjusted composite scores and age (in years) was examined. Analyses of variance (ANOVAs) were then performed to examine other demographic associations with performance across each age-adjusted composite measure. Also, age-adjusted composite scores were examined for association with health and school functioning variables. Effect sizes are reported as Cohen's *d*, with cutoffs of .20, .50, and .80 indicating small, medium, and large effects, respectively.

Similarly, normalized composite scaled scores were constructed for the gold standard measures, and associations between corresponding NIH Toolbox CFB and gold standard composite scores were computed. These had to be done separately for younger (ages 3 to 6 years) and older (8 to 15 years) because the gold standard fluid measures were different for

these age groups. Finally, again for the separate age groups, other psychometric properties of the NIH Toolbox CFB and gold standard composite scores were compared.

RESULTS

Test-retest Reliability

For the 66 participants across all ages (3 to 15 years) who were retested, excellent test-retest correlations were observed: *r*'s = .92, .95, and .96 for Toolbox Crystallized, Fluid, and Total composite scores, respectively; all *df* = 64 and *p*'s < .0001. The total longitudinal sample sizes for the two separate age groups arguably are too small to confidently generate reliability estimates. However, the 3- to 6-year-old subgroup with longitudinal data (*n* = 38) also evidenced good test-retest correlations on the full Toolbox composite scores (*r*'s = .74, .86, and .89 for Toolbox Crystallized, Fluid, and Total, *df* = 36, *p*'s < .0001). Their reliability estimates for the Short Fluid and Short Total composite scores were also acceptable (*r*'s = .78 and .73, *df* = 36, *p* < .0001), although reliability of the Short Crystallized score (Picture Vocabulary only) was more modest (*r*(36) = .50, *p* = .002). The 28 participants aged 8 to 15 years who were retested obtained robust reliability estimates on the full composite scores, although these were somewhat lower than those seen with the total child group (*r*'s = .85, .76, and .88 for Crystallized, Fluid, and Total, *df* = 26, *p*'s < .0001).

"Practice effects" were computed as the means and standard deviations of the difference between the follow-up composite scaled score and the baseline composite scaled score, with significance of the effect being tested with *t*-tests for dependent means. For the total child group (ages 3 to 15 years, *n* = 66), the Toolbox Crystallized Composite evidenced virtually no practice effect over an average two week test-retest interval: *mean practice effect* = 0.00, *SD* = 1.18, *t*(65) = −0.03, *p* = .98. However, the Toolbox Fluid Composite score showed a significant effect of about a half of a scaled score point (*mean* = 0.50, *SD* = .96, *t*(65) = 4.27, *ES=0.52, p* < .0001), and the Toolbox Total Composite also had a modest practice effect (*mean* = 0.27, *SD* = .87, *t*(65) = 2.57, *ES=0.31, p* = .01). Interestingly, degree of practice effect was not significantly correlated with age for any of the Toolbox composite scores (Crystallized *r* = .14, Fluid *r* = .06, Total *r* = .16, *df* = 64, all *p*'s > .19).

Age Effects

Pearson correlation coefficients were used to examine the effects of age with neuropsychological test performance on the Toolbox composite measures. As shown in Figure 1, there was clear evidence that the Toolbox composite measures are sensitive to developmental growth during childhood. Across the 3- to 15-year age span (*N* = 208), age was highly correlated with performance on the Toolbox Crystallized Composite (*r*(203) = .87), Toolbox Fluid Composite (*r*(205) = .86), and Toolbox Total Composite (*r*(205) = .88) (all *p*'s < .0001). Furthermore, Figure 1 shows almost overlapping, linear effects of age on the Toolbox Crystallized and Fluid Composite scores across the full age span from 3 to 15 years. It is noted as well that in this total child group the correlation between Crystallized and Fluid Composite scores is similarly strong (*r*(203) = .89, *p* < .0001). There is some evidence that these two composite scores are beginning to "decouple" slightly in the older

group (r = .64 in the 8- to 15-year-old group versus r = .77 in the 3- to 6-year-old group), but this difference was not significant (*Fisher's z*, p = .07).

Other Demographic Differences

When adjusted for age, there were no significant "effects" of gender, mother's education, or family income on the total group's Toolbox Crystallized Cognition Composite, Toolbox Fluid Cognition Composite, or Toolbox Total Composite. There were statistically significant ethnicity effects, albeit with small effect sizes (*ES*'s = .25, .28, .29), on the age-adjusted Toolbox Crystallized Cognition Composite ($F(2,192)$ = 4.96, p = .008), Fluid Cognition Composite ($F(2,194)$ = 5.93, p = .003), and Total Composite (F(2,194) = 7.80, p = .0006). On each of the composite scores, Caucasian children scored higher than African American children.

Relations with Health Status and School Performance

Again using the total participant sample (ages 3–15), children's reported health status was significantly related (but with small effect sizes) to the age-adjusted Toolbox Crystallized Cognition Composite ($F(1,199)$ = 7.62, p = .006; *ES* = .21), Fluid Cognition Composite ($F(1,199)$ = 4.03, p = .046; *ES* = .15) and Total Composite ($F(1,199)$ = 8.41, p = .004; *ES* = . 20). In each case the children described as having "excellent" health performed somewhat better than those described as having less than excellent (poor to very good) health.

Associations of Toolbox composite scores with reported school performance were assessed only in the older (school age, 8–15) children, because many of the younger children were not yet in formal school settings. In these, school performance was strongly associated (medium to large effect sizes) with age-adjusted scores on the Toolbox Crystallized Cognition Composite ($F(1,77)$ = 34.48, p < .0001; *ES* = .86), Fluid Cognition Composite ($F(1,77)$ = 13.48, p = .0004; *ES* = .65), and Total Composite ($F(1,77)$ = 34.72, p < .0001; *ES* = .85). Children who were reported to have "above average" school performance (vs. average or below average) scored consistently higher on all composite scores. Similarly, children who were reported to have needed "special" (remedial) classes or tutoring in school performed worse on age-adjusted Toolbox Crystallized ($F(1,75)$ = 10.53, p = .002; *ES* =.67), Fluid ($F(1,75)$ = 3.25, p = .075; *ES* = .44), and Total ($F(1,75)$ = 12.03, p = .0009; *ES* = .72) composite scores.

Construct Validity

Convergent—As noted above, comparisons of results on the Toolbox and gold standard composite scores required separate analyses for the two age groups (3 to 6 years; 8 to 15 years) because gold standard measures were different at different ages (see also Chapter 1). In Table 1 are correlations between analogous Toolbox and gold standard composite scores for the two age groups. These results show good convergent validity for the full and short composite scores in both age groups: median correlations of .88 for Crystallized, .70 for Fluid, and .88 for Total.

Discriminant—Only modest evidence for discriminant validity is provided by slightly lower correlations between Toolbox Crystallized and Gold Standard Fluid Composite scores

(median $r = .71$) and between Toolbox Fluid and Gold Standard Crystallized Composite scores ($r = .72$). These latter results should be considered in light of the very high correlation between the Toolbox Crystallized and Fluid Composite scores themselves ($r = .89$ for the full sample).

Evidence of discriminant validity for the short battery composites (for children with ages of 3–6 years only) consisted of somewhat lower correlations for non-analogous than those for analogous composites: the correlation between Toolbox Crystallized and Gold Standard Fluid Short Composite scores was .48 ($df = 107$, $p < .0001$), and that between Toolbox Fluid and Gold Standard Crystallized Short Composite scores was .67 ($df = 112$, $p < .0001$).

Developmental trajectories of Toolbox versus Gold Standard composites— Given the results in Table 1, it should not be surprising that the Toolbox and Gold Standard composite scores show very similar developmental trajectories (age effects) in childhood (see Figures 2a and 2b). Although there appears to be a sharp increase in the slope in Figure 2b after the age of 10, this probably is an artifact based upon the need to include more than one age in the last two categories (because of small sample sizes). Also similar to results for the Toolbox measures presented above, there were no significant gender effects on age-adjusted gold standard composite scores for either age group. Maternal education was related to gold standard composite scores only in the 3- to 6-year-old group (all p's < .05). Also only in the younger group, family income was associated with just one gold standard composite (Crystallized; $p < .05$).

As noted above, significant ethnicity effects were seen on all Toolbox composite scores when the full age range (3 to 15 years) was considered, reflecting somewhat better cognitive performance in the Caucasian than in the African American children. However, when the smaller age subgroups were analyzed separately, most of these differences became nonsignificant trends (exceptions were Toolbox Fluid and Total Composite scores in the 3- to 6-year-olds). Similarly, in these age subgroups the ethnicity group differences in the gold standard composite scores showed mostly nonsignificant trends (exceptions were Gold Standard Fluid Composite scores in the 3- to 6-year-olds, and Gold Standard Crystallized Composite scores in the 8- to 15-year-olds).

As was reported above in relation to the Toolbox composite scores, strong and consistent relations were found between all gold standard composite scores and reported school functioning in the older, school-age group. This was true in relation to parental reports of children's overall school performance (above average vs. below average or average) as well as reports about the needs for special classes or tutoring.

Discussion

The results from the validation study suggest that the proposed Toolbox Crystallized Cognition, Fluid Cognition, and Cognitive Function (i.e., total) Composite scores provided reliable measures of important aspects of cognition for children between the ages of 3 and 15. Although the subgroup of our participants that was reassessed in this study was relatively small ($n = 66$), the test-retest reliability estimates on the Toolbox composite scores

(r's = .92 to .96) are comparable to those seen with well-established cognitive summary scores in the literature (e.g., IQ scores on the Wechsler intelligence scales; Sattler, 2001; WAIS-III WMS-III Technical Manual, 1997).

The longitudinal data in this study also indicate a modest practice effect for the Toolbox Fluid Cognition Composite but not the Toolbox Crystallized Cognition Composite. This is expected, because of the types of abilities that are reflected in these composite scores. Tests of "fluid" cognition involve new learning and adapting to novel stimuli and task requirements; when such tests are repeated the examinee tends to show improved performance because the test stimuli and required tasks are more familiar (less novel). This benefit of prior test exposure is not as likely when previously learned knowledge and skills (crystallized cognition) are being assessed. Other examples of this fluid versus crystallized difference include repeated administrations of intelligence test batteries, where Perceptual Organization and Processing Speed (fluid) composite scores show more improvement than "crystallized" measures of Verbal Comprehension (WAIS-III WMS-III Technical Manual, 1997). On the other hand, even the largest practice effect observed for a Toolbox composite was rather modest (average of only about ½ scaled score point on the Toolbox Fluid Cognition Composite, with a medium effect size of 0.54). Nevertheless, longitudinal studies that use NIH Toolbox Fluid Cognition Composite or any other fluid cognition measures must control for practice effects to avoid misinterpreting these as "real" improvements due to development or some form of intervention (medical or educational).

The cross-sectional, developmental data in this study are based upon a much larger sample of normal children (n = 208). Across the age span represented in this sample (3 to 15 years), all three NIH Toolbox CFB composite scores demonstrated a strong, linear developmental trajectory. Two points about this are worth noting. First, whereas in adults fluid and crystallized abilities tend to diverge with older age because of greater "normal aging" effects on fluid cognition (e.g., Heaton et al., 2003), during childhood the developmental effects on the two types of cognition show parallel positive trends (Sattler, 2001). Indeed, in Figure 1 the mean age trajectories of the Toolbox Crystallized and Fluid Cognition Composite scores are virtually overlapping. Furthermore, to our knowledge the NIH Toolbox CFB is unique in its ability to track with the same instruments both developmental effects and effects of adult aging across the entire lifespan (ages 3 to 85 years).

In this chapter we also examined possible relations between multiple demographic factors and children's performance on the NIH Toolbox CFB. By far, children's age was the most powerful demographic predictor. Once age was corrected, there were no significant effects of gender, or SES as indexed by mother's education level or family income. Consistent with typical findings in adult groups (Heaton et al., 2003; Heaton et al., 2004; Heaton, Ryan & Grant, 2009; Norman et al, 2011), Caucasian children performed somewhat better than their African American counterparts on the Toolbox composite scores. The effect sizes of these differences were "small," however, and lower than is typically seen in studies of adult cognition. Nevertheless, when normative standards are being used to evaluate possible developmental or acquired cognitive _disorders_, failure to adjust for even small demographic effects can increase classification errors (e.g., Heaton et al., 2003).

Although the current study focused on a rather restricted (12-year) age range in childhood, we were unable to find comparable gold standard tests that allowed us to explore convergent and discriminant validity across these ages. For this reason, NIH Toolbox CFB versus gold standard associations and comparisons had to be done separately for the 3- to 6-year-olds and the 8- to 15-year-olds. The analyses provided evidence for good convergent validity in both age groups. There was less support for discriminant validity when comparing Toolbox Fluid and Crystallized Cognition Composite scores with Gold Standard Crystallized and Fluid Composite scores, respectively. This is not surprising, given the exceptionally strong, positive age effects on all cognitive measures in this study (both Toolbox and gold standard), as well as robust correlations between the Toolbox Fluid and Crystallized Cognition Composite scores across the full age span (3 to 15 years).

There may be situations where researchers choose to use the single Toolbox Cognitive Function Composite as a measure of overall cognitive ability in children. Although there were strong correlations between the Toolbox Fluid and Crystallized Cognition Composite scores, it is unlikely that this collection of neuropsychological tests is simply measuring one underlying ability or cognitive factor. There continues to be much debate about the nature and development of fluid and crystallized abilities and general intelligence (see Blair, 2006). Research on newer measures of children's fluid cognition is needed, particularly examining how they relate to the development of crystallized abilities and the underlying brain maturation associated with the development of these skills. The NIH Toolbox CFB is currently being utilized in a large NIH-funded multi-site study of brain development, and appears to hold promise in this regard.

Our assessments of children's health status and academic performance were limited to rather gross indicators by maternal reports. The health status ratings, in particular, lacked specificity, and their small (but statistically significant) effects on the age-corrected NIH Toolbox CFB composite scores probably were due to particular health problems that may have influenced cognitive development and/or educational experiences in some children. Future research should be directed at validating the CFB measures for detecting effects on cognition of specific health conditions of interest (e.g., trauma or infections involving the brain, developmental disorders, other chronically disabling conditions).

Perhaps this study's most impressive evidence of the NIH Toolbox CFB Composite scores' criterion validity (other than the convergent validation with gold standard measures) was their strong associations with maternal ratings of children's school performance and needs for special classes or tutoring (large effect sizes). This might not be surprising for the Toolbox Crystallized Cognition Composite, because this composite reflects previously acquired semantic knowledge (vocabulary) and reading skills that are both products of more or less successful educational experiences. Although the Toolbox Fluid Cognition Composite does not directly reflect what the child has learned in the past, it does indicate the status of other cognitive abilities that may be considered necessary to succeed in future schoolwork (i.e., attention, working memory, processing speed, episodic memory, and executive function). These findings suggest, therefore, that the CFB composite scores may have validity for assessing aspects of children's cognition that are important for educational success.

In our experience, it is sometimes more difficult to administer the entire NIH Toolbox CFB to children under age 7 years than to school age children. Younger children may require more time to become familiar with the task instructions and may fatigue more easily, despite the relatively short time required to administer the CFB. Examination of potential "short form" scores for children ages 3 to 6 years was promising. The use of three scores from two CFB tests to create the Short Toolbox Fluid Cognitive score produced results that were very similar to the Toolbox Fluid Composite score (which includes six scores from five tests). However, reliability of the Vocabulary measure (as a single measure of crystallized ability) was much lower than expected in the youngest children. Conceivably this will improve when the test goes to national norming in late 2011. The computer response format used with the validation sample was a touch screen and it was noted that some of the youngest children had difficulty with the touch screen. In the norming version the touch screen has been replaced with two button keys (arrow keys) which will likely reduce unintended response errors. The national norming study will involve a larger, more representative sample of both children and adults, and is expected to provide further confirmation of the reliability and validity of the Crystallized, Fluid and Total Composite scores introduced here.

References

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

Blair C. How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. Behavioral and Brain Sciences. 2006; 29(2):109–25. discussion 125–60. 10.1017/S0140525X06009034 [PubMed: 16606477]

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

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. 4. Circle Pines, MN: American Guidance Services; 2007. (PPVT-4)

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.; 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. 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]

Korkman, M.; Kirk, U.; Kemp, S. NEPSY. 2. San Antonio, TX: Harcourt Assessment; 2007. (NEPSY-II)

Li SC, 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– 63.10.1111/j.0956-7976.2004.01503003.x [PubMed: 15016286]

Norman MA, Moore DJ, Taylor M, Franklin D, Cysique L, Ake C, Lazarretto D, Vaida F, 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 EPub May 2011, 4, 1, 11.

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 Preschool and Primary Scale of Intelligence. 3. San Antonio, TX: The Psychological Corporation; 2002. (WPPSI-III)

Wechsler, D. Wechsler Intelligence Scale for Children. 4. San Antonio: The Psychological Corporation; 2003.

Weintraub, S.; Dikman, SS.; Heaton, RK.; Tulsky, DS.; Zelazo, PD.; Bauer, PJ.; Carlozzi, NE.; Slotkin, J.; Blitz, D.; Wallner-Allen, K.; Fox, NA.; Beaumont, JL.; Mungas, D.; Nowinski, CJ.; Richler, J.; Deocampo, JA.; Anderson, JE.; Manly, JJ.; Borosh, B.; Havlik, R.; Gershon, R. NIH Toolbox for the Assessment of Behavioral and Neurological Function: Cognition Domain Instruments. (submitted)Manuscript submitted for publication

Wilkinson, GS.; Robertson, GJ. Wide Range Achievement Test 4 professional manual. Lutz, FL: Psychological Assessment Resources; 2006.
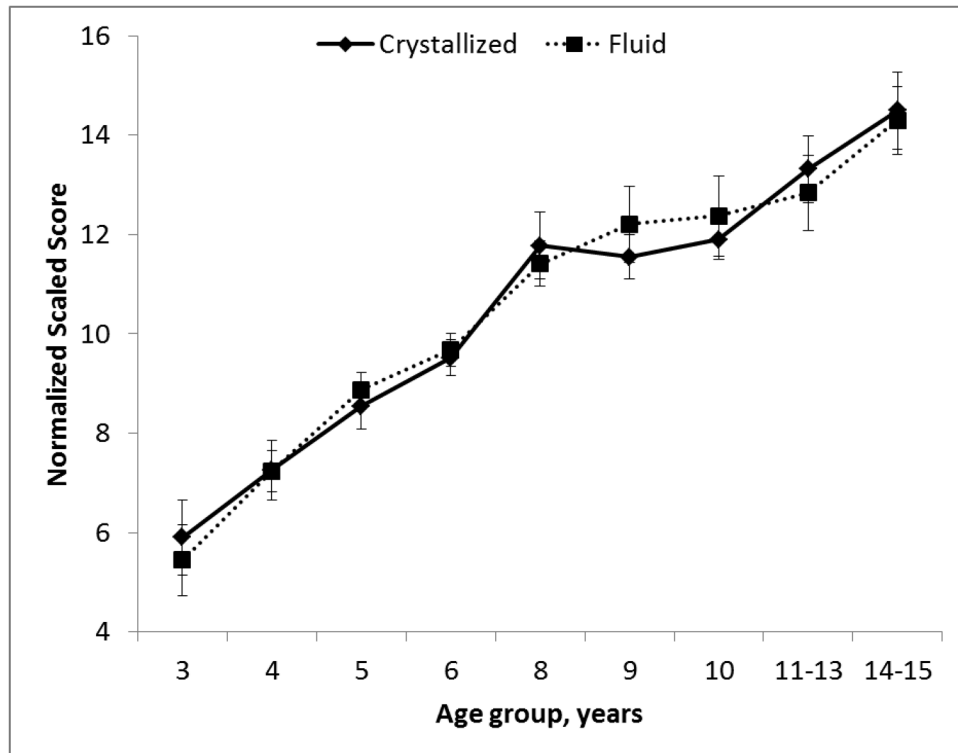
Figure 1.
Performance on the Toolbox Crystallized Cognition Composite and the Toolbox Fluid Cognition Composite across age groups. Error bars represent +/− 2 SE.
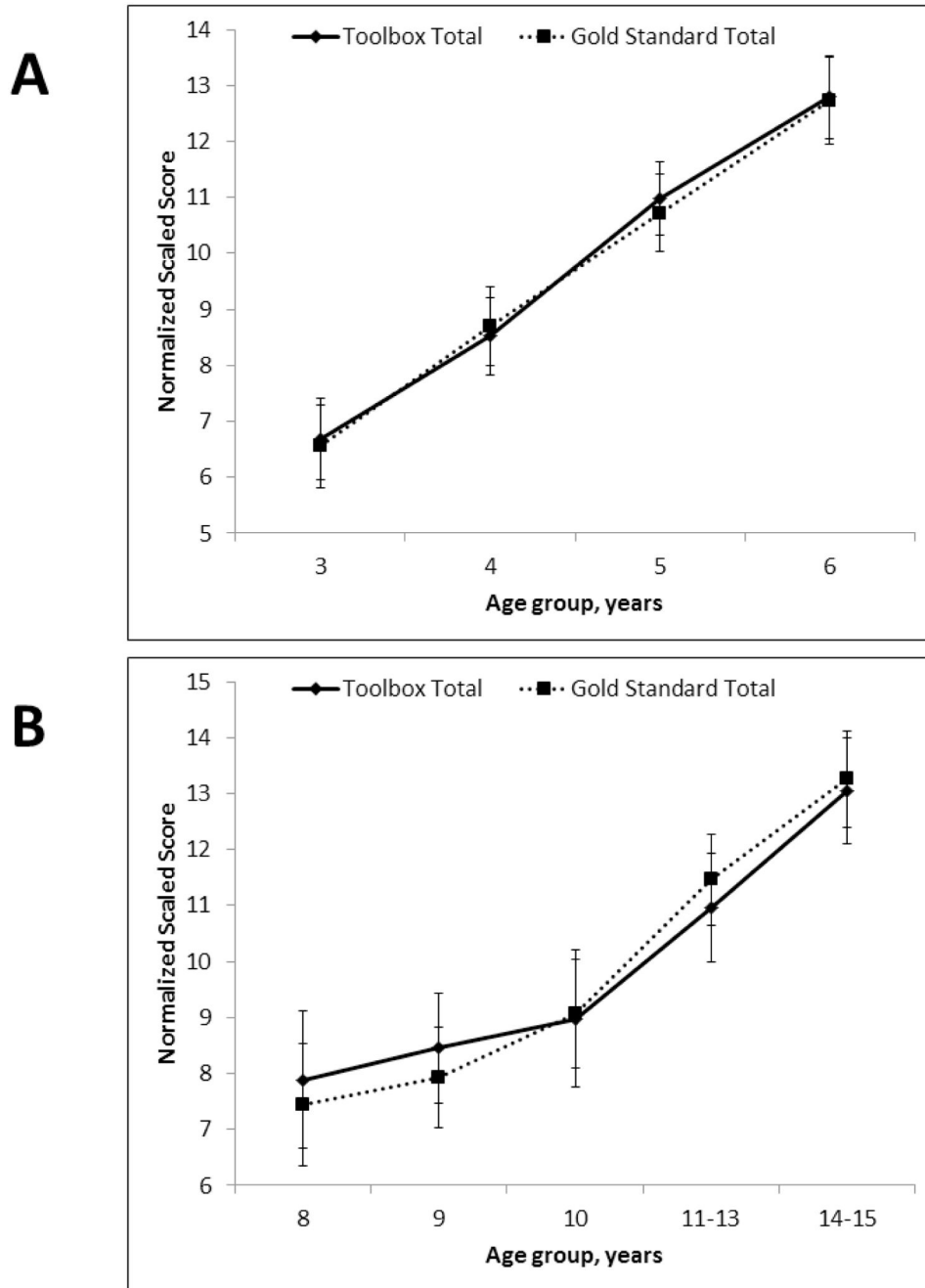
Figure 2.
Performance on the Toolbox Cognitive Function Composite (Total) and the Gold Standard Total Composite across ages, plotted separately for younger (3 to 6 years) and older (8 to 15 years) children. Error bars represent +/− 2 SE.

Table 1

Correlations between full and short NIH Toolbox CFB and gold standard composite scores for younger (3 to 6 years) and older (8 to 15 years) children (all *p*'s <. 0001)

	Full Composites		Short Composites
	Age 3–6	**Age 8–15**	**Age 3–6**
Crystallized	.88 (df=114)	.90 (df=85)	.73 (df=109)
Fluid	.78 (df=111)	.70 (df=85)	.69 (df=111)
Total	.90 (df=114)	.88 (df=85)	.80 (df=113)