HHS Public Access

Author manuscript

*J Int Neuropsychol Soc*. Author manuscript; available in PMC 2015 October 12.

Published in final edited form as:

*J Int Neuropsychol Soc*. 2014 July ; 20(6): 620–629. doi:10.1017/S1355617714000472.

# NIH Toolbox Cognition Battery (CB): Validation of Executive Function Measures in Adults

**Philip David Zelazo**[1], **Jacob E. Anderson**[1], **Jennifer Richler**[2], **Kathleen Wallner-Allen**[3], **Jennifer L. Beaumont**[4], **Kevin P. Conway**[5], **Richard Gershon**[4], and **Sandra Weintraub**[4]

[1]University of Minnesota, Minneapolis, Minnesota [2]Indiana University, Bloomington, Indiana [3]Westat, Inc., Rockville, Maryland [4]Northwestern University, Evanston, Illinois [5]National Institutes of Health, Bethesda, Maryland

## Abstract

This study describes psychometric properties of the NIH Toolbox Cognition Battery (NIHTB-CB) executive function measures 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 two computer-based instruments assessing executive function: the Dimensional Change Card Sort (a measure of cognitive flexibility) and a flanker task (a measure of inhibitory control and selective attention). Participants completed the NIHTB-CB, corresponding gold standard convergent and discriminant measures, and sociodemographic questionnaires. A subset of participants ($N$ = 89) was retested 7 to 21 days later. Results reveal excellent sensitivity to age-related changes during adulthood, excellent test–retest reliability, and adequate to good convergent and discriminant validity. The NIH Toolbox EF measures can be used effectively in epidemiologic and clinical studies.

### Keywords

Cognitive control; Cognitive flexibility; Inhibitory control; Lifespan development; Standardized testing; Validation

## NIH TOOLBOX COGNITION BATTERY (CB): MEASURING EXECUTIVE FUNCTION AND ATTENTION

Executive function (EF), also called cognitive control, is a construct that encompasses the top-down neurocognitive processes involved in the conscious, goal-directed control of thought, action, and emotion—processes such as cognitive flexibility, inhibitory control, and working memory (Miyake et al., 2000). These processes depend on the integrity of neural networks involving prefrontal cortex (PFC), the anterior cingulate cortex, and other regions

Correspondence and reprint requests to: Philip David Zelazo, Institute of Child Development, University of Minnesota, 51 East River Road, Minneapolis, MN 55455-0345. zelazo@umn.edu.

(see Miller & Cohen, 2001; Zelazo & Lee, 2010, for review), and they are required for solving problems flexibly, attending selectively and ignoring distractions, and keeping information in mind.

Two measures were designed to assess executive function (EF) as part of the National Institutes of Health (NIH) Toolbox for Neurological and Behavioral Function-Cognition Battery (CB). The NIH Toolbox provides a set of standardized tests that can be used longitudinally across the lifespan, from ages 3 to 85 years. In this article, we describe the creation of the EF measures and report data on the psychometric properties of the tests, including sensitivity to age-related differences, test/retest reliability, and construct validity, for participants ages 20–85. Pediatric data are reported separately (see Zelazo et al., 2013). We conclude with a discussion of the limitations of the battery as well as the implications of the EF measures for the study of lifespan cognitive development.

## Executive Function across the Lifespan

Developmental research on EF has revealed that when considered across the lifespan, EF appears to follow an inverted-U-shaped curve, rising and then falling (see Jacques & Marcovitch, 2010, for a review). It should be noted, however, that the absence of measures suitable across a wide range of childhood ages and into adulthood has made it difficult to characterize the lifespan development of EF in detail. For example, it has been difficult to compare levels of EF seen in elderly adults with those seen in early childhood. Whereas most childhood measures of EF are too easy for elderly adults, many classic neuropsychological measures of EF, such as the Wisconsin Card Sorting Test (Grant & Berg, 1948) or the Color-Word Stroop task (Stroop, 1935), are either too difficult for young children or inappropriate for other reasons (e.g., the Stroop task assumes not only that participants be literate, but also that reading be fully automatized).

Nonetheless, lifespan studies from middle childhood into late adulthood have typically revealed both the rise and the fall of EF (Bialystok & Craik, 2010; Jacques & Marcovitch, 2010; Salthouse & Davis, 2006). For example, research on cognitive flexibility has used a variety of task switching paradigms and generally found decreases in switch costs (i.e., the extent to which participants slow down on, or in the context of, switch trials) across childhood and adolescence, and subsequent increases in older adults (Cepeda, Kramer, & Gonzalez de Sather, 2001; Crone, Bunge, Van der Molen, & Ridderinkhof, 2006; Reimers & Maylor, 2005; Zelazo, Craik, & Booth, 2004). Adults and children over the age of approximately 6 years generally slow down to maintain a high level of accuracy, so individual differences in EF are often manifested as degrees of slowing rather than as errors (e.g., Davidson, Amso, Cruess-Anderson, & Diamond, 2006). In a cross-sectional study of participants ages 7 to 82 years, for example, Cepeda et al. (2001) found that switch costs (above and beyond perceptual speed, working memory, and non-switch reaction time) decreased from childhood into young adulthood and then stayed fairly constant until approximately 60 years of age, after which they increased. Similarly, Zelazo and colleagues (2004) found that the number of perseverative errors on a different measure of task switching, a version of the Dimensional Change Card Sort (DCCS, see below), decreased from childhood (*M* age = 8.8 years) to young adulthood (22.3 years) and then increased

again in late adulthood (71.1 years). Theoretical accounts of the changes in EF occurring during adulthood have emphasized changes in a range of underlying processes, from attentional resources (e.g., Craik & Byrd, 1982) to processing speed (e.g., Salthouse, 1996) to inhibition (e.g., Dempster, 1992; Hasher & Zacks, 1988).

The inverted U-shaped pattern of EF development fits with what is known about age-related changes in PFC structure and function across the lifespan (Raz, 2000; Zelazo & Lee, 2010), and it is consistent with research suggesting that neural circuits involved in more complex cognitive functions, such as EF, may be especially vulnerable to disruption, due to decreases in gray matter volume (e.g., Sowell et al., 2003), neuronal shrinking (e.g., Terry, De Teresa, & Hansen, 1987), reduction in the length of myelinated axons (e.g., Marner, Nyengaard, Tang, & Pakkenberg, 2003), and other processes. Various well-known “signs” of aging, such as increased forgetting and unwanted intrusions of irrelevant material into one’s speech, may be attributable, to some extent, to impaired EF associated with the aging of prefrontal cortex (e.g., von Hippel, 2007).

## Toolbox Measurement

To provide an assessment of EF across the lifespan, the NIH Toolbox Cognition Battery was designed to include measures of the three aspects of EF identified in Miyake et al.’ s (2000) tripartite model, including cognitive flexibility and inhibitory control, as well as a measure of working memory that will be described separately (see Tulsky et al., this issue). Working memory is considered separately because although it is an aspect of EF, it is often studied on its own, or as one of several forms of memory. For the NIH Toolbox Cognition Battery, one measure each of cognitive flexibility and inhibitory control was identified that (a) was freely available (in the public domain) and (b) had the potential to be modified, in an iterative fashion, to meet the usability objectives of the NIH Toolbox—namely, that they be computer-administered, very brief (<5 min), relatively immune to practice effects, and suitable for participants between the ages 3 and 85 years. These measures were then subjected to an iterative process of measure development that involved modifying existing measures in order to meet these objectives and to satisfy the criteria of four NIH Toolbox working groups: geriatric, pediatric, accessibility, and cultural sensitivity (see Weintraub et al., this issue). For example, the measures were modified so that the instructions were easy to understand and the visual displays were engaging for participants at all ages. The font sizes, image sizes, types of motoric response required, and colors of stimuli (with respect to color blindness) were all designed to increase the accessibility of the measures for the general U.S. population, including the oldest participants. The number of trials in each task was minimized while maximizing test-retest reliability and validity.

## Executive Function-Cognitive Flexibility

The Dimensional Change Card Sort (DCCS) was selected as the measure of cognitive flexibility, also known as task switching or set shifting. This task, designed by Zelazo and colleagues (e.g., Frye, Zelazo, & Palfai, 1995; Zelazo, 2006), is based on Luria’s seminal work on rule use and has been used extensively to study the development of EF in childhood. In the standard version of the DCCS for children, individuals are shown two target cards (e.g., a blue rabbit and a red boat) and asked to sort a series of bivalent test cards

(e.g., red rabbits and blue boats) first according to one dimension (e.g., color), and then according to the other (e.g., shape). Most 3-year-olds perseverate during the post-switch phase, continuing to sort test cards by the first dimension, whereas most 5-year-olds switch flexibly (e.g., Dick, Overton, & Kovacs, 2005; Kirkham, Cruess, & Diamond, 2003; Zelazo, Müller, Frye, & Marcovitch, 2003). More challenging versions of this task have been used with older children, adolescents, and young and old adults (e.g., Diamond & Kirkham, 2005; Morton, Bosma, & Ansari, 2009; Zelazo et al., 2004; see Zelazo, 2006). Both the standard version of this task and a more challenging version have shown excellent test–retest reliability in childhood (*ICCs* = .90–.94; Beck, Schaefer, Pang, & Carlson, 2011).

### Executive Function-Inhibitory Control and Attention

A version of the Eriksen flanker task (Eriksen & Eriksen, 1974) was adapted from the Attention Network Test (ANT; e.g., Rueda et al., 2004; Fan, McCandliss, Sommer, Raz, & Posner, 2002). In a flanker task, participants are required to indicate the left–right orientation of a centrally presented stimulus while inhibiting attention to the potentially incongruent stimuli that surround it (i.e., the flankers, typically two on either side). In the traditional flanker task, the stimuli are arrows pointing left or right, whereas in the ANT version used with children, the stimuli are fish (designed to be more engaging and also larger, which makes the task easier). The version created for the NIH Toolbox Cognition Battery includes both an easier fish block and a more difficult arrows block. On some trials, the orientation of the flanking stimuli is congruent with the orientation of the central stimulus, and on others it is incongruent. Performance on the incongruent trials provides a measure of inhibitory control in the context of visual selective attention (which can also be considered a measure of executive attention; e.g., Fan et al., 2002), and shows clinical utility in identifying deficits associated with neurological disorder (e.g., Coubard et al., 2011).

To assess the construct validity of the new EF measures for adult participants, we examined data from a validation study of the NIH Toolbox Cognition Battery and compared performance on the NIH Toolbox DCCS Test and the NIH Toolbox Flanker Inhibitory Control and Attention Test to performance on an established measure of EF (for a measure of convergent validity) and on an established measure of a different construct, receptive vocabulary (for a measure of discriminant validity). An initial publication (Weintraub et al., 2013) introduced the Cognition Battery along with the rest of the NIH Toolbox and provided an overview and summary data from the entire validation sample, including both children and adults. The second set of publications comprised a monograph focused on the data from children ages 3–15 years (Zelazo & Bauer, 2013). The current report is the first detailed presentation of the validation data from adults (20–85 years). In addition to convergent and discriminant validity, we examined sensitivity to age-related changes in performance across adulthood, and a random subset of participants was retested after ~ 2 weeks, allowing us to measure test–retest reliability and practice effects.

# METHOD

## Participants

Demographic information about the participants (including study sites) in the validation study is described in detail by Weintraub et al. (this issue). Proper consents were obtained and approved by the relevant Institutional Review Boards. Briefly, there were 268 adults ages 20 to 85 years in this sample, recruited through a registry of healthy older individuals ($N$ = 62) maintained at the Northwestern University Cognitive Neurology and Alzheimer's Disease Center (CNADC) in Chicago, IL, and through community flyers around four university-based testing sites (in Evanston, IL, $N$ = 25; West Orange, NJ, $N$ = 92, Seattle, WA, $N$ = 67, and Chicago, IL, $N$ = 12). The flyers advertised for healthy volunteers but no further health screening or exclusions were applied. Stratified sampling guidelines were used to enhance demographic variability, and the final sample was indeed distributed across gender (119 males), highest education level (Mean = 13.4 years; range = 4–20; $SD$ = 2.9), and race/ethnicity (148 non-Hispanic White, 75 Black or African American, 38 Hispanic or Latino, and 7 multi-racial). Education was further categorized as less than high school graduate (25%), high school graduate or some college (37%), and Bachelor's degree or higher (38%). As described below, data from some participants were missing or excluded from the final analyses (e.g., for failing to reach criterion during practice trials), leaving final samples that ranged from $n$ = 237 to $n$ = 264 for each measure. To assess test–retest reliability, 89 participants (approximately 33% of the adult sample) were randomly selected to be retested after 7 to 21 days (*mean interval* = 15.5 days, $SD$ = 4.8).

## Measures

**NIH Toolbox Dimensional Change Card Sort (DCCS) Test—**This measure consisted of four blocks (practice, pre-switch, post-switch, and mixed) that were presented on a touch-screen monitor. Instructions appeared visually on the monitor and were also read aloud by the experimenter to all participants.

During the practice block, participants were given a series of practice trials on which they were instructed to sort a bivalent test stimulus (either a green rabbit or a white boat) by either shape or color. The test stimulus was presented on a central screen and participants sorted it by touching one of two laterally presented target stimuli (white rabbit and green boat). The initial dimension (shape or color) by which participants sorted was counterbalanced across participants. A response was recorded when participants touched either of the target stimuli, though subsequent research suggests that a simple key press (i.e., using keys that are spatially congruent with the target stimuli) works equally well at all ages; accordingly, the key press will be employed in newer versions of NIH Toolbox instruments. See Figure 1 for trial specifications, although it should be noted that different stimuli were used for practice, and during practice only, participants were also given a feedback screen that indicated whether or not their response was correct.

Participants were required to sort 3 out of 4 practice items correctly in order to proceed, and if they did not meet this criterion, they could receive up to two additional series of 4 practice trials (i.e., they were given as many as 3 chances to meet the criterion). Once the criterion

was met for the first sorting dimension, participants were trained on the second dimension. If a participant failed to meet the criterion for either dimension, the task was stopped. No participants failed to meet the practice criterion.

When the practice criterion was met for both dimensions, participants were administered test trials. The trial structure for test trials was the same as for practice trials (see Figure 1), although different shapes (ball/truck) and colors (yellow/ blue) were used in the test trials, and no feedback was provided. First, a pre-switch block of 5 trials was administered in which participants needed to sort by the same dimension (e.g., color) that was used in the immediately preceding practice block. If participants sorted correctly on 4 of 5 trials, they were told to switch to sorting by the other dimension (e.g., shape), and 5 post-switch trials were administered. If participants failed to reach the criterion on either the pre- or post-switch block, the test was terminated. Participants who met the criterion for post-switch trials were informed that they would now be asked to switch back and forth between dimensions and were given 50 mixed trials, including 40 “dominant” and 10 “non-dominant” trials presented in a pseudorandom order (with 2–5 dominant trials preceding each non-dominant trial). The dominant dimension was always the sorting dimension used in the post-switch block (e.g., shape).

As described elsewhere (Zelazo et al., 2013), the NIH Toolbox DCCS was scored using a new two-vector scoring method combining both accuracy and, for participants who maintained a high level of accuracy (> 80% correct), reaction time (RT) into one score. (It is also possible, using the NIH Toolbox, to examine RT and accuracy data separately.) On tasks like the NIH Toolbox DCCS, older children and adults have a tendency to slow down (> RT) to maintain a high level of accuracy, and this RT slowing provides an index of EF “cost” (Davidson et al., 2006). In contrast, younger children (below approximately 6 years of age) usually do not show a speed/accuracy trade-off, but instead continue to respond quickly at the expense of accuracy. For these participants, accuracy provides a better index of EF cost. This new scoring method includes children and adults on the same metric scale, allowing for instrument comparisons across the lifespan.

Performance was scored based on the total number of test trials completed, whether these included only the pre-switch block, both pre- and post-switch blocks, or all blocks. For all participants who received the mixed block, however, the number of included trials was truncated to the first 30 because preliminary analyses indicated increased variability in performance toward the end of the task, including effects that may interact with age, and because our objective was to create measures that were as brief as possible while still maintaining reliability. For all participants, accuracy was considered first, and scored on a scale from 0 to 5. Participants were given 0.125 points (5 points divided by 40 total task trials: 5 pre-, 5 post-, and 30 mixed-block trials) for every correct response they made on trials they received. Expressed as an equation:

$$Accuracy\ Score = 0.125 * Number\ Correct\ Responses \quad \text{(Equation1)}$$

For participants who were accurate on 80% or fewer trials, final scores were equal to accuracy scores. For those who were accurate on more than 80% of trials, an RT score was

also calculated based on each participant's median RT on correct non-dominant trials from the mixed block. First, RTs lower than 100 milliseconds (ms) or greater than 3 standard deviations (*SDs*) from each participant's mean RT were discarded as outliers because these trials were unlikely to provide a valid measure of performance. Second, median RTs were calculated. Third, because RTs typically have a positively skewed distribution, a log (Base 10) transformation was applied to each participant's median RT score to create a more normal distribution of scores.

Based on the distribution of scores in the validation data, the minimum median RT for scoring was set to 500ms and the maximum to 3000ms. Median RTs that fell outside of this range but within the range of 100 to 10,000ms were truncated for the purposes of RT score calculation so that RTs between 100 and 500ms were set equal to 500 ms and RTs between 3000ms and 10,000ms were set equal to 3000ms. This truncation did not introduce any ceiling or floor effects. Log values were algebraically rescaled from a log(500) – log(3000) range to a 0–5 range. Rescaled scores were reversed such that smaller RT log values were at the upper end of the 0–5 range whereas larger RT log values were at the lower end. Once the rescaled RT scores were obtained, they were added to the accuracy scores for participants who achieved the accuracy criterion of greater than 80%.

**NIH Toolbox Flanker Inhibitory Control and Attention Test—**The flanker task consisted of a practice block, a fish block, and an arrows block. During practice sessions, which administered fish stimuli, participants were instructed to press one of two laterally presented arrow "buttons" on the touch screen, each corresponding to the direction a middle fish was pointing (see Figure 2 for the trial structure and the timing of each stimulus). For all trials, the word "middle" was presented visually on each trial, to remind participants to attend to the middle stimulus. Participants were given 4 trials (2 congruent and 2 incongruent) and were required to respond correctly on at least 3 out of the 4 to advance to the test trials. If they did not meet this criterion, they could receive up to two additional series of practice trials. No participants failed to meet the practice criterion. Testing was terminated if the participant failed to meet criterion by the third practice trial set. Participants who passed the practice block received a block of 25 fish trials (16 congruent and 9 incongruent trials) presented in a pseudorandom order (with 1–3 congruent trials preceding each incongruent trial). Participants who responded correctly to 5 or more of the 9 incongruent trials proceeded to the arrows block. In the arrows block, the stimuli consisted of arrows instead of fish, but the structure of this block was otherwise identical to the fish block (25 trials, with 16 congruent and 9 incongruent).

Scores for the NIH Toolbox Flanker were created using a procedure analogous to the one used for scoring the NIH Toolbox DCCS. That is, a two-vector method incorporated accuracy and, for participants who maintained a high level of accuracy (> 80% correct), RT. Again, preliminary analyses indicated increased variability of performance during the last few trials of the test, so scoring was based on the first 20 (out of 25) trials in each block (fish/arrows). Accuracy and RT vector scores were calculated using the same formulae used for the NIH Toolbox DCCS, and each type of score ranged from 0 to 5. That is, Equation 1 was used to determine accuracy scores (based on both congruent and incongruent trials), and

RT data were scored in the same way as in the NIH Toolbox DCCS. RT scores were added to the accuracy scores for participants who achieved an accuracy level of 80% or better.

## Validation Measures

**Convergent validity measure—**Convergent validity was estimated by measuring the relation between the two Toolbox EF measures, and between each Toolbox EF measure and Color-Word Interference Inhibition raw scores from the Delis-Kaplan Executive Function Scales (D-KEFS; Delis, Kaplan, & Kramer, 2001). The D-KEFS Color-Word Test, which is based on the Stroop Color-Word test (Stroop, 1935), measures both cognitive flexibility and the ability to inhibit attention and responding. First, participants name the colors of color patches (red, green, or blue). Next, they read the names of the colors ("red," "green," "blue") that appear in black print. Finally, they are shown color words printed in non-corresponding colors (e.g., "red" printed in blue ink) and told to ignore the printed words and report only the color in which each word is printed. (The test also includes an Interference-switching condition that was not administered). Color-Word Interference Inhibition raw scores provide a measure of flexibility and susceptibility to interference; higher scores indicate better performance. This measure was selected by consensus among the cognition domain team as the "gold standard" measure that most closely corresponded in test format (speeded) and in targeted constructs (inhibitory control and cognitive flexibility) to the aspects of EF assessed by the NIH Toolbox. The Toolbox EF measures also served as convergent validation measures for each other, and because these measures were expected to correlate highly, only one measure of convergent validity was used for both measures of EF.

**Discriminant validity measure—**The Peabody Picture Vocabulary Test, 4th Edition (PPVT-4; Dunn & Dunn, 2007) was used as a discriminant measure, and this measure was also selected by consensus by the cognition domain team as the "gold standard" measure that most clearly assessed a construct that is distinct from EF. On each trial, an array of four pictures was provided along with a word describing one of the pictures. Participants were asked to point to, or say the number of, the picture that best corresponds to the word. The test was administered and scored using the standard protocol. As a test of receptive vocabulary, the PPVT-4 provides an index of relatively crystallized knowledge, and is often used as a proxy for full scale IQ or general developmental level (Kline, 2001). The PPVT-4 was selected because it has good psychometric properties across the entire age range for the battery (ages 3 to 85 years), allowing for use of the same metric for all ages.

## Data Analysis

For the NIH Toolbox DCCS Test, 20 adults did not complete the task, data from 2 adults were judged to be invalid based on the examiner's notes (e.g., stopping mid task, etc.), and 2 adults had too few correct trials upon which to base RT scores (i.e., fewer than 2 correct trials), leaving a final sample for this measure of $n = 244$. An additional 11 adults were flagged as outliers (but not excluded from the final sample) based on scores that were less than 4 points—well below the range of scores for the remainder of the adult sample (see below). For the NIH Toolbox Flanker, 27 adults did not complete the task, data from 2 adults were judged to be invalid based on the examiner's notes, and 2 adults had too few correct trials upon which to base RT scores (i.e., fewer than two correct trials), leaving a

final sample for this measure of $n = 237$. An additional 4 adults were flagged as outliers based on scores that were less than 4 points (see below). Analyses were conducted both with and without the participants flagged as outliers; results were similar, and results based on the larger sample size (with outliers) are reported here.

For both the NIH Toolbox measures and the validation measures, scaled scores were created by first ranking the raw scores of all participants between the ages of 20 and 85 years, then applying a normative transformation to the ranks to create a standard normal distribution, and finally rescaling the distribution to have a mean of 10 and a standard deviation of 3. These scaled scores were used in all analyses and not adjusted for age. Both Pearson and intraclass correlation coefficients (ICC) with 95% confidence intervals were calculated to evaluate test–retest reliability. ICC less than .4 indicated poor test–retest reliability, .4–.75 adequate, and .75 or higher good to very good. Practice effects were evaluated using paired *t* tests and effect sizes (mean change from time 1 to time 2 / *SD* of Time 1) were calculated as a standardized estimate of the mean change (Cohen, 1992). Convergent validity was assessed by examining the correlations between the two NIH Toolbox measures and between these measures and the D-KEFS Inhibition score. Convergent validity correlations less than .3 were considered poor, .3–.6 adequate, and .6 or higher good to very good evidence of convergent validity. Evidence of discriminant validity consisted of lower correlations with selected "gold standard" measures of a *different* cognitive construct. Discriminant validity was assessed by examining correlations with the PPVT-4. Analyses of variance (ANOVAs) were then performed to examine other demographic associations with performance, adjusted for age and other relevant covariates. Effect sizes are reported as Cohen's *d*, with cutoffs of .20, .50, and .80 indicating small, medium, and large effects, respectively.

## RESULTS

### Age Effects

Figure 3 presents performance (scaled scores) on the new NIH Toolbox measures and the two established measures as a function of age across the entire adult sample. Performance on the EF measures appears to peak in the mid-20s followed by a gradual decline through 85 years. Exploratory correlations were performed to describe the relations between EF and age in years around the observed peak (see Table 1). Across ages 20–29 years, age was positively related to scores on the NIH Toolbox DCCS, but not the NIH Toolbox Flanker. All EF measures were significantly negatively related to age between 25 and 85 years. The relations between age and PPVT-4 scores were as follows: for ages 20–29 years, $r = .14$; for ages 30–85 years, $r = .10$.

### Test–Retest Reliability and Practice Effects

As shown in Table 1, the new NIH Toolbox measures showed excellent test–retest reliability (see Table 1 for both ICCs and Pearson correlations). The established measures also showed excellent reliability. Significant practice effects ($p < .001$), with medium effects sizes, were observed over the 2-week test–retest interval for the NIH Toolbox DCCS (mean practice effect in scaled score units = .95; $SD = 1.61$; $d = .33$), NIH Toolbox Flanker (mean practice

effect = .79; *SD* = 1.73; *d* = .27), and D-KEFS Inhibition (mean practice effect = .79; *SD* = 1.38; *d* = .27), although not for the PPVT-4 (mean practice effect = .17; *SD* = 1.20; *d* = .06).

### Construct Validity

**Convergent validity—**D-KEFS Inhibition raw scores were positively correlated with scores on the NIH Toolbox DCCS ($r(237) = .55$; $p<.0001$) and the NIH Toolbox Flanker ($r(229) = .52$; $p< .0001$). Performance on the NIH Toolbox DCCS was positively correlated with performance on the NIH Toolbox Flanker ($r(226) = .71$; $p < .0001$). Together, these results indicate adequate to good convergent validity.

**Discriminant validity—**Compared to the convergent validity correlations, much lower correlations were observed between performance on the PPVT-4 and scores on the NIH Toolbox DCCS, $r(242) = .06$; $p = .37$, and the NIH Toolbox Flanker, $r(234) = .06$; $p = .35$), indicating good discriminant validity.

### Other Demographic Factors

Table 2 shows effect sizes for other demographic factors known to influence cognitive performance, after adjusting for age and other additional relevant demographic variables. There was no significant difference between males and females for either NIH Toolbox EF measure. White participants scored higher on both measures than Black participants (small effect size) and Hispanic participants (medium effect size). Finally, college graduates scored higher than those with less than a high school education and those whose highest level was high school. Effect sizes were medium in both cases. For both measures, participants whose highest level was high school scored higher than those with less than a high school education (large effect size).

## DISCUSSION

As part of the NIH Toolbox, new versions of the DCCS and a flanker task were created to provide brief assessments of two aspects of EF: cognitive flexibility (NIH Toolbox DCCS) and inhibitory control in the context of visual selective attention (NIH Toolbox Flanker Inhibitory Control and Attention). Together with a separate measure of working memory (see Tulsky et al., this issue), these measures capture the three aspects of EF identified by Miyake et al. (2000) in their factor-analytic work with adults.

Overall, results reveal excellent sensitivity to age-related changes during adulthood, excellent test–retest reliability, and adequate to good convergent and discriminant validity. Repeated administration (with a 1–3 week lag) revealed practice effects that were comparable for the NIH Toolbox measures and the established measure of convergent validity. Consistent with previous studies, our findings suggest that EF peaks in early adulthood and then declines gradually with age. Performance on the NIH Toolbox DCCS continued to improve between the ages of 20 and 29 years, but this was not the case on the NIH Toolbox Flanker. Future research with multiple measures per construct will be needed to determine whether cognitive flexibility and inhibitory control follow a different developmental trajectory during early adulthood.

Performance on the new NIH Toolbox measures of EF was positively correlated with performance on an established measure of EF, the D-KEFS Color Word Test, and they were correlated with each other. In contrast, the new measures were not related to receptive vocabulary as measured by the PPVT-4, which may be considered a proxy for general intellectual level. Together with the results from the pediatric data from this validation study (Zelazo et al., 2013), which showed substantial age-related increases in discriminant validity during childhood, these findings support the suggestion that EF becomes increasingly differentiated from general intellectual function over the course of childhood and adolescence. This pattern is also consistent with the suggestion made by Cepeda et al. (2001) that some of the processes underlying EF might become more differentiated with age, perhaps because, as previous studies have shown, nonexecutive abilities such as vocabulary are relatively spared from age-related decline (see Craik & Salthouse, 2000).

Results from this sample revealed significant associations between performance on the NIH Toolbox measures of EF and self-reported race/ethnicity and highest educational level obtained. The effect sizes for race/ethnicity were small to medium, whereas the effect sizes for educational level were medium to high. This pattern of results is generally consistent with the existing literature, which has revealed similar correlations between socioeconomic status and EF in childhood (e.g., Noble, Norman, and Farah, 2004). Further research is required to understand the nature of these demographic associations, but there is growing evidence that prefrontally mediated skills, which modulate many other neural functions, may be especially vulnerable to disruption given their dependence on the integrity of these other neural functions, their protracted developmental course, and their sensitivity to the effects of stress (e.g., Masten et al., 2012).

Having methodologically sound measurement tools that can be used over a broad age span will be of considerable value to the field for a number of reasons, from basic to applied. On the basic side, the new NIH Toolbox EF measures will facilitate efforts to describe the development of EF across the lifespan. On the more applied side, being able to follow a much wider range of developmental pathways longitudinally will be useful for intervention studies, and it may provide opportunities to identify key mechanisms underlying important developmental outcomes, including physical and mental health and academic and social success (e.g., Moffit et al., 2011). The new NIH Toolbox EF measures are the first standardized measures of EF appropriate for use from early childhood (age 3 years) to old age (age 85).

EF and attention have emerged as major foci of research in part because they predict a wide range of important developmental outcomes. Moffitt et al. (2011), for example, found that EF measured in childhood predicts (as a gradient) physical health, substance dependence, socioeconomic status, and the likelihood of a criminal conviction at age 32 years, even after controlling for social class of origin and IQ. In addition, numerous developmental disorders are characterized by deficits in EF, which suggests that its development is fragile and easily disrupted. EF deficits emerging in adulthood are key signs of neurocognitive decline (e.g., von Hippel, 2007). At the same time, however, there is growing evidence that EF and attention are malleable—even in adulthood (e.g., Olesen, Westerberg, & Klingberg, 2003). The far-reaching consequences of EF and attention underscore the importance of a complete

understanding of their developmental course across the lifespan, and the NIH Toolbox Cognition Battery represents an important step toward achieving this goal.

## LIMITATIONS AND FUTURE DIRECTIONS

The creation of the NIH Toolbox Cognition Battery is an important advance in the study of cognitive function and its lifespan development, and its availability promises to accelerate discoveries through use of common methods across disparate laboratories and even disciplines. It does, however, have its limitations. First, given the many aspects of neurological and behavioral function assessed by the NIH Toolbox, a complete and comprehensive assessment of validity was not possible. The use of a single measure (D-KEFS Color Word Interference: Inhibition score) as an index of convergent validity for both Toolbox EF measures was less than ideal. Further research, using a wider range of convergent measures of EF (i.e., multiple convergent measures of cognitive flexibility and inhibitory control), is needed to assess more fully the validity of the Toolbox measures.

Second, the Toolbox measures of EF focus on its relatively “cool” cognitive aspects, often associated with lateral prefrontal cortex and elicited by relatively abstract, decontextualized problems. Additional measures would be needed to assess the more “hot” aspects of executive function, which are more associated with the orbitofrontal cortex and seen in situations that are emotionally and motivationally significant because they involve meaningful rewards or punishers (e.g., Happaney, Zelazo, & Stuss, 2004).

Further testing of the NIH Toolbox is now underway with a larger number of children, adolescents, and adults to establish national (U.S.) norms for performance. As part of this research, norms will also be provided for a Spanish-language version of the NIH Toolbox. A direction for future research will be to examine the utility of the NIH Toolbox for children and adults suffering from neurological insult or injury or neurocognitive developmental disorders. Although the current validation study included a diverse range of participants, it was not designed specifically to evaluate cognitive health in individuals with neurocognitive disabilities.

## ACKNOWLEDGMENTS

This study was 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. *Disclaimer:* The views and opinions expressed in this article are those of the authors and do not necessarily represent the views of the National Institute on Drug Abuse, the National Institutes of Health, or any other governmental agency. *COI Disclosures: Dr. Zelazo* serves on the editorial boards of Child Development, Development and Psychopathology, Frontiers in Human Neuroscience, Cognitive Development, Emotion, and Developmental Cognitive Neuroscience, and Monographs of the Society for Research in Child Development. He is a Senior Fellow of the Mind and Life Institute and President of the Jean Piaget Society. He receives research funding from the Canadian Institute for Health Research (Grant # 201963), NIDDK/NICHD (1699-662-6312), and the Character Lab. *Mr. Anderson* reports no disclosures. *Dr. Richler* is funded by NIH/NCRR grant UL1RR025761. *Dr. Wallner-Allen* reports no disclosures. *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. *Dr. Conway* 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. *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.

## REFERENCES


Beck DM, Schaefer C, Pang K, Carlson SM. Executive function in preschool children: Test-retest reliability. Journal of Cognition and Development. 2011; 12:169–193. [PubMed: 21643523]

Bialystok, E.; Craik, FIM. Structure and process in life-span cognitive development. In: Overton, WF., editor. Cognition, biology, and methods across the lifespan. Volume 1 of the Handbook of life-span development. Hoboken, NJ: Wiley; 2010. p. 195-225.

Cepeda NJ, Kramer AF, Gonzalez de Sather JCM. Changes in executive control across the life span: Examination of task-switching performance. Developmental Psychology. 2001; 37:715–730. [PubMed: 11552766]

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

Coubard OA, Ferrufino L, Boura M, Gripon A, Renaud M, Bherer L. Attentional control in normal aging and Alzheimer's disease. Neuropsychology. 2011; 25:353–367. [PubMed: 21417533]

Craik, FIM.; Byrd, M. Aging and cognitive deficits: The role of attentional resources. In: Craik, FIM.; Trehub, S., editors. Aging and cognitive processes. New York: Plenum Press; 1982. p. 191-211.

Craik, FIM.; Salthouse, TA., editors. The handbook of aging and cognition. 2nd ed.. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.

Crone EA, Bunge SA, Van der Molen MW, Ridderinkhof KR. Switching between tasks and responses: A developmental study. Developmental Science. 2006; 9:278–287. [PubMed: 16669798]

Davidson MC, Amso D, Cruess-Anderson L, Diamond A. Development of cognitive control and executive functions from 4–13 years: Evidence from manipulations of memory, inhibition and task switching. Neuropsychologia. 2006; 44:2037–2078. [PubMed: 16580701]

Delis, DC.; Kaplan, E.; Kramer, JH. Delis-Kaplan executive function system. San Antonio, TX: Pearson (The Psychological Corporation); 2001.

Dempster FN. The rise and fall of the inhibitory mechanism: Toward a unified theory of cognitive development and aging. Developmental Review. 1992; 12:45–75.

Diamond A, Kirkham N. Not quite as grown-up as we like to think. Psychological Science. 2005; 16:291–297. [PubMed: 15828976]

Dick AS, Overton WF, Kovacs SL. The development of symbolic coordination: Representation of imagined objects, executive function, and theory of mind. Journal of Cognition and Development. 2005; 6:133–161.

Dunn, LM.; Dunn, DM. Peabody Picture Vocabulary Test. 4th edition. San Antonio, TX: Pearson; 2007.

Eriksen BA, Eriksen CW. Effects of noise letters upon the identification of a target letter in a nonsearch task. Perception and Psychophysics. 1974; 16:143–149.

Fan J, McCandliss BD, Sommer T, Raz A, Posner MI. Testing the efficiency and independence of attentional networks. Journal of Cognitive Neuroscience. 2002; 14:340–347. [PubMed: 11970796]

Frye D, Zelazo PD, Palfai T. Theory of mind and rule-based reasoning. Cognitive Development. 1995; 10:483–527.

Grant DA, Berg EA. A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type-card-sorting problem. Journal of Experimental Psychology. 1948; 38:404–411. [PubMed: 18874598]

Happaney K, Zelazo PD, Stuss DT. Development of orbitofrontal function: Current themes and future directions. Brain and Cognition. 2004; 55:1–10. [PubMed: 15134839]

Hasher, L.; Zacks, RT. Working memory, comprehension, and aging: A review and new view. In: Bower, GH., editor. The psychology of learning and motivation: Advances in research and theory. Vol. 22. New York: Academic Press; 1988. p. 193-225.

Jacques, S.; Marcovitch, S. Development of executive function across the lifespan. In: Overton, WF., editor. Cognition, biology, and methods across the lifespan. Volume 1 of the Handbook of life-span development. Hoboken, NJ: Wiley; 2010. p. 431-466.

Kirkham N, Cruess LM, Diamond A. Helping children apply their knowledge to their behavior on a dimension-switching task. Developmental Science. 2003; 6:449–467.

Kline, RB. Brief cognitive assessment of children: Review of instruments and recommendations for best practice. In: Andrews, JJW.; Sakloske, DH.; Janzem, HL., editors. Handbook of psychoeducational assessment. San Diego, CA: Academic Press; 2001. p. 103-132.

Marner L, Nyengaard JR, Tang Y, Pakkenberg B. Marked loss of myelinated nerve fibers in the human brain with age. Journal of Comparative Neurology. 2003; 462:144–152. [PubMed: 12794739]

Masten AS, Herbers JE, Desjardins CD, Cutuli JJ, McCormick CM, Sapienze JK, Zelazo PD. Executive function skills and school success in young children experiencing homelessness. Educational Researcher. 2012; 41:375–384.

Miller EK, Cohen JD. An integrative theory of prefrontal cortex function. Annual Review of Neuroscience. 2001; 24:167–202.

Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD. The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. Cognitive Psychology. 2000; 41:49–100. [PubMed: 10945922]

Moffit TE, Arsenault L, Belsky D, Dickson N, Hancox RJ, Harrington H, Caspi A. A gradient of childhood self-control predicts health, wealth, and public safety. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108:2693–2698. [PubMed: 21262822]

Morton JB, Bosma R, Ansari D. Age-related changes in brain activation associated with dimension shifts of attention: An fMRI study. Neuroimage. 2009; 46:249–256. [PubMed: 19457388]

Noble KG, Norman MF, Farah MJ. Neurocognitive correlates of socioeconomic status in kindergarten children. Developmental Science. 2004; 8(1):74–87. [PubMed: 15647068]

Olesen PJ, Westerberg H, Klingberg T. Increased prefrontal and parietal activity after training of working memory. Nature Neuroscience. 2003; 7:75–79. [PubMed: 14699419]

Raz, N. Aging of the brain and its impact on cognitive performance: Integration of structural and functional findings. In: Craik, FIM.; Salthouse, TA., editors. The handbook of aging and cognition. 2nd ed.. Mahwah, NJ: Lawrence Erlbaum Associates; 2000. p. 1-90.

Reimers S, Maylor EA. Task switching across the lifespan: Effects of age on general and specific switch costs. Developmental Psychology. 2005; 41:661–671. [PubMed: 16060812]

Rueda MR, Fan J, McCandliss BD, Halparin JD, Gruber DB, Lercari LP, Posner MI. Development of attentional networks in childhood. Neuropsychologia. 2004; 42:1029–1040. [PubMed: 15093142]

Salthouse TA. The processing-speed theory of adult age differences in cognition. Psychological Review. 1996; 103:403–428. [PubMed: 8759042]

Salthouse TA, Davies HP. Organization of cognitive abilities and neuropsychological variables across the lifespan. Developmental Review. 2006; 26:31–54.

Sowell ER, Peterson BS, Thompson PM, Welcome SE, Henkenius AL, Toga AW. Mapping cortical change across the human life span. Nature Neuroscience. 2003; 6:309–315. [PubMed: 12548289]

Stroop JR. Studies of interference in serial verbal reactions. Journal of Experimental Psychology. 1935; 18:643–662.

Terry RD, De Teresa R, Hansen LA. Neocortical cell counts in normal adult aging. Annals of Neurology. 1987; 21:530–539. [PubMed: 3606042]

von Hippel W. Aging, executive functioning, and social control. Current Directions in Psychological Science. 2007; 16:240–244.

Weintraub S, Dikmen SS, Heaton RK, Tulsky DS, Zelazo PD, Bauer PJ, Gershon RC. Cognition assessment using the NIH Toolbox. Neurology. 2013; 80(11 Suppl. 3):S54–S64. [PubMed: 23479546]

Zelazo PD. The Dimensional Change Card Sort: A method of assessing executive function in children. Nature Protocols. 2006; 1:297–301. [PubMed: 17406248]

Zelazo, PD.; Anderson, JE.; Richler, J.; Wallner-Allen, K.; Beaumont, JL.; Weintraub, S. NIH Toolbox Cognitive Function Battery (CFB): Measuring executive function and attention. 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 Monographs of the Society for Research in Child Development. Vol. 78. 2013. p. 16-33.

Zelazo PD, Bauer PJ. National Institutes of Health Toolbox—Cognitive Function Battery (NIH Toolbox CFB): Validation for children between 3 and 15 years. Monographs of the Society for Research in Child Development. 2013; 78(4)

Zelazo PD, Craik FIM, Booth L. Executive function across the life span. Acta Psychologica. 2004; 115:167–184. [PubMed: 14962399]

Zelazo, PD.; Lee, WSC. Brain development: An overview. In: Overton, WF., editor. Cognition, biology, and methods across the lifespan. Volume 1 of the handbook of lifespan development. Hoboken, NJ: Wiley; 2010. p. 89-114.

Zelazo PD, Müller U, Frye D, Marcovitch S. The development of executive function in early childhood. Monographs of the Society for Research in Child Development. 2003; 68(3)

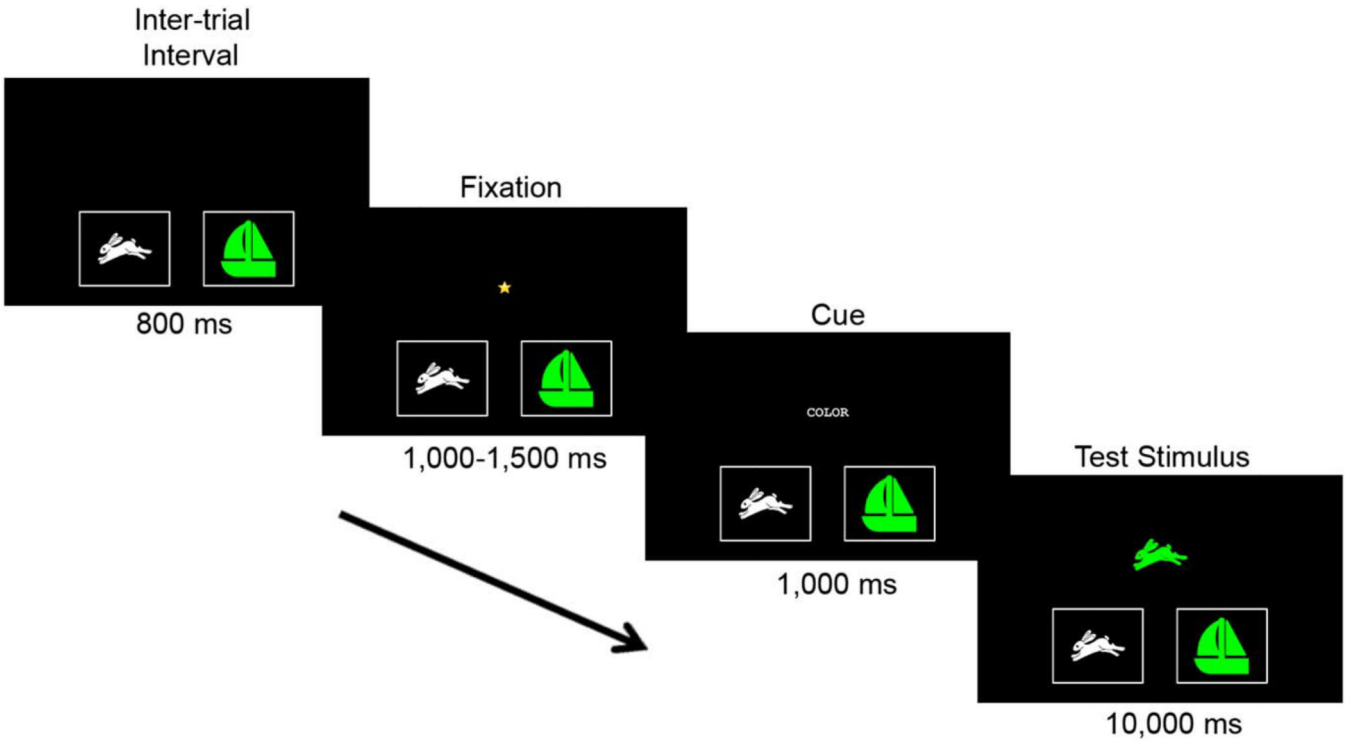


**Fig. 1.**

Trial sequence for the NIH Toolbox Dimensional Change Card Sort Test (with practice stimuli). All NIH Toolbox-related materials are ©2012 Northwestern University and the National Institutes of Health.

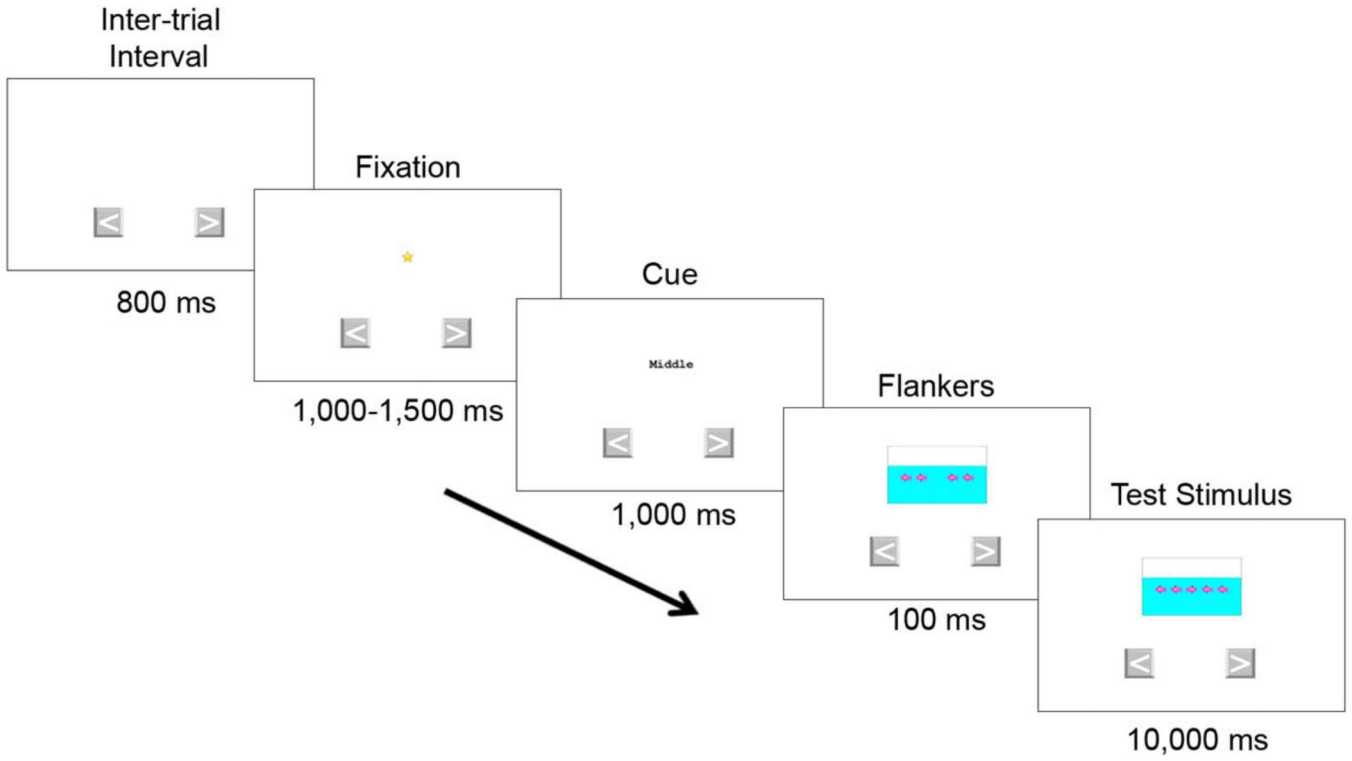


**Fig. 2.**

Trial sequence for the NIH Toolbox Flanker Inhibitory Control and Attention Test (fish block). All NIH Toolbox-related materials are ©2012 Northwestern University and the National Institutes of Health.

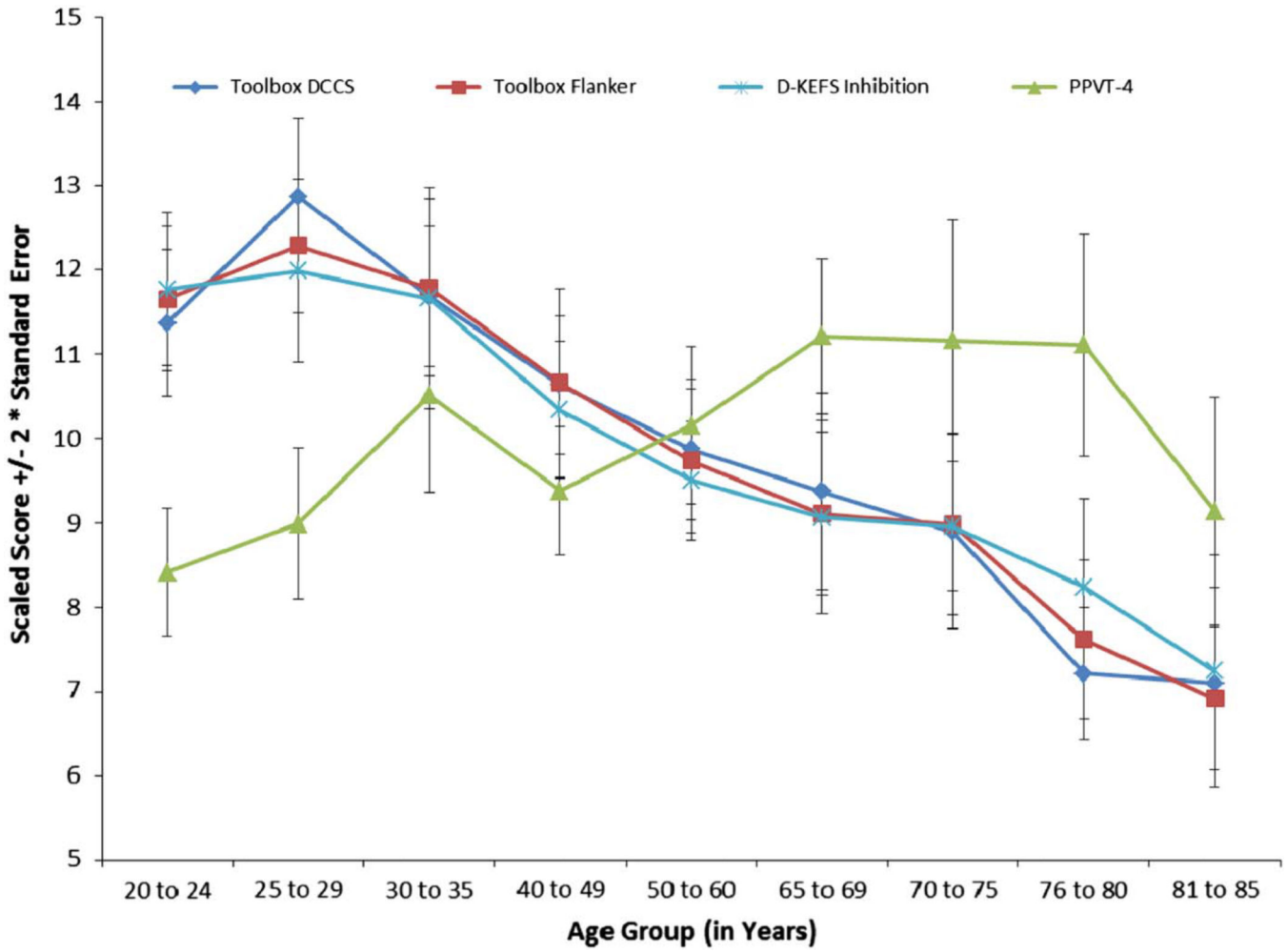


**Fig. 3.**

Performance on the NIH Toolbox DCCS, the Toolbox Flanker, and the measures of convergent (D-KEFS Color Word Inhibition Raw Score) and discriminant validity (PPVT-4), as a function of age.

## Table 1

Pearson correlations with age & test-retest Pearson correlations (r) and intraclass correlation coefficients (ICC) for toolbox measures of EF and measures of convergent (D-KEFS Inhibition) and discriminant validity (PPVT-4)

| Measure | Age Correlation: Entire Sample (*n*) | Age Correlation: 20–29 years (*n*) | Age Correlation: 25–85 years (*n*) | Test-retest *r*, *ICC* (*n*, *95%* CI) |
|---|---|---|---|---|
| Toolbox DCCS | −.55** (244) | .31* (60) | −.59** (208) | .85, .81 (78, .72–.87)** |
| Toolbox Flanker | −.54** (237) | .10 (60) | −.55** (201) | .85, .83 (73, .74–.89)** |
| D–KEFS: Inhibition | −.48** (257) | .06 (62) | −.47** (221) | .90, .87 (88, .81–.91)** |
| PPVT-4 | .24** (263) | .14 (62) | .16 (227) | .92, .92 (89, .88–.95)** |

* $p < .05$,

** $p < .0001$.

**Table 2**

Effect sizes (ES) for comparisons of scores between groups

| | Toolbox DCCS | Toolbox flanker |
|---|---|---|
| *ES* (male vs. female)[1] | −.10 | −.13 |
| *P* | .339 | .246 |
| *ES* (Black vs. white)[2] | −.23 | −.38 |
| *ES* (Hispanic vs. white)[2] | −.54 | −.53 |
| *ANOVA p* | .003 | <.001 |
| *ES* (college vs. < high school)[3] | .39 | .44 |
| *ES* (college vs. high school grad)[3] | .29 | .34 |
| *ES* (high school grad vs. < high school)[3] | .91 | .90 |
| *ANOVA p* | .018 | .007 |

[1] Gender comparison adjusted for age and education.

[2] Race/ethnicity comparisons adjusted for gender, age, and education. ‘White’ = non-Hispanic White; ‘Black’ = Black or African American; and ‘Hispanic’ = Hispanic or Latino.

[3] Education comparison adjusted for age. ‘College’ = Bachelor’s degree or higher; ‘High School Grad’ = high school graduate or some college; ‘< High School’ = less than completion of high school.