Published in final edited form as:
Neuropsychology. 2014 January ; 28(1): 1–10. doi:10.1037/neu0000001.

The NIH Toolbox Cognition Battery: Results from a Large Normative Developmental Sample (PING)

Natacha Akshoomoff[1,2], **Erik Newman**[1,2], **Wesley K. Thompson**[1,3], **Connor McCabe**[2], **Cinnamon S. Bloss**[4], **Linda Chang**[5], **David G. Amaral**[6], **B. J. Casey**[7], **Thomas M. Ernst**[5], **Jean A. Frazier**[8], **Jeffrey R. Gruen**[9], **Walter E. Kaufmann**[10], **Tal Kenet**[11], **David N. Kennedy**[8], **Ondrej Libiger**[4], **Stewart Mostofsky**[10], **Sarah S. Murray**[4], **Elizabeth R. Sowell**[12], **Nicholas Schork**[4], **Anders M. Dale**[13,14], **Terry L. Jernigan**[1,2,14,15], and **for the Pediatric Imaging, Neurocognition, and Genetics Study**

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

[2]Center for Human Development, University of California, San Diego, La Jolla, CA

[3]Stein Institute for Research on Aging, University of California, San Diego, La Jolla, CA

[4]Scripps Genomic Medicine, Scripps Translational Science Institute and Scripps Health, La Jolla, CA

[5]Department of Medicine, University of Hawaii and Queen's Medical Center, Honolulu, HI

[6]Department of Psychiatry and Behavioral Sciences and The M.I.N.D. Institute, University of California, Davis, Sacramento, CA

[7]Sackler Institute for Developmental Psychobiology, Weil Cornell Medical College, New York, NY

[8]Department of Psychiatry, University of Massachusetts Medical School, Boston, MA

[9]Departments of Pediatrics and Genetics, Yale University School of Medicine, New Haven, CT

[10]Kennedy Krieger Institute and Johns Hopkins University School of Medicine, Baltimore, MD

[11]Department of Neurology and Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA

[12]Department of Pediatrics, University of Southern California, Los Angeles, CA and Children's Hospital, Los Angeles, CA

[13]Department of Neurosciences, University of California, San Diego, La Jolla, CA

[14]Department of Radiology, University of California, San Diego, La Jolla, CA

[15]Department of Cognitive Science, University of California, San Diego, La Jolla, CA

Abstract

Corresponding Author: Natacha Akshoomoff, Ph.D., UC San Diego Department of Psychiatry, School of Medicine and Center for Human Development, 9500 Gilman Drive, La Jolla, CA 92093-0115, Telephone: 858-822-2757, FAX: 858-822-1602, nakshoomoff@ucsd.edu.
Walter E. Kaufmann is now in the Department of Neurology, Boston Children's Hospital and Harvard Medical School, Connor McCabe is now in the Department of Psychology, University of Washington, and Sarah S. Murray is now in the Department of Pathology, University of California, San Diego. Elizabeth R. Sowell was previously at the University of California, Los Angeles, which was also a PING data collection site.

PING data are disseminated by the PING Coordinating Center at the Center for Human Development, University of California, San Diego. Anders M. Dale is a founder of and holds equity interest in CorTechs Labs, La Jolla, CA, and serves on its scientific advisory board. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

Objective—The NIH Toolbox Cognition Battery (NTCB) was designed to provide a brief, efficient computerized test of key neuropsychological functions appropriate for use in children as young as 3 years of age. This report describes the performance of a large group of typically developing children and adolescents and examines the impact of age and sociocultural variables on test performance.

Method—The NTCB was administered to a sample of 1020 typically developing males and females ranging in age from 3 to 20 years, diverse in terms of socioeconomic status (SES) and race/ethnicity, as part of the new publicly accessible Pediatric Imaging, Neurocognition, and Genetics (PING) data resource, at 9 sites across the United States.

Results—General additive models of nonlinear age-functions were estimated from age-differences in test performance on the 8 NTCB subtests while controlling for family SES and genetic ancestry factors (GAFs). Age accounted for the majority of the variance across all NTCB scores, with additional significant contributions of gender on some measures, and of SES and race/ethnicity (GAFs) on all. After adjusting for age and gender, SES and GAFs explained a substantial proportion of the remaining unexplained variance in Picture Vocabulary scores.

Conclusions—The results highlight the sensitivity to developmental effects and efficiency of this new computerized assessment battery for neurodevelopmental research. Limitations are observed in the form of some ceiling effects in older children, some floor effects, particularly on executive function tests in the youngest participants, and evidence for variable measurement sensitivity to cultural/socioeconomic factors.

Keywords

Computerized Assessment; Cognitive Development; Socioeconomic Status

The NIH Toolbox Cognition Battery: Results from a Large Normative Developmental Sample

The NIH Toolbox for Assessment of Neurological and Behavioral Function[SM] was commissioned by the NIH Blueprint for Neuroscience Research to provide brief, efficient, and highly accessible tests to measure cognitive and emotional health, and provide a "common currency" for neurological research (Gershon et al., 2010). The NIH Toolbox divides tests into four domain batteries: Cognition, Sensation, Motor, and Emotion. The NIH Toolbox Cognition Battery (NTCB) was designed to tap key functions (executive function, attention, episodic memory, working memory, language, and processing speed) across the lifespan (ages 3 to 85 years). This computerized approach provides an economical method for assessing a wide range of cognitive abilities, which is appealing for large-scale studies. For pediatric studies, this also has the advantage of providing the same set of measures for use with young children, older children, and adolescents in an appealing format that provides automated data collection, storage, and scoring.

Here we describe the age-related changes in performance on the NTCB from a large normative sample of 1000+ individuals ranging in age from 3 to 20 years. These data were collected as part of the Pediatric Imaging, Neurocognition, and Genetics (PING) Study. PING is a data resource that also includes highly standardized and carefully curated MRI data and genome-wide single nucleotide polymorphism (SNP) data, as shown in recent publications (Bakken et al., 2012; Brown et al., 2012; Fjell et al., 2012; Walhovd et al., 2012). By openly sharing these data, PING will allow researchers to examine links between the development of cognitive functions, genetic variation, and patterns of brain structure and connectivity in typically developing individuals. This normative dataset can also be

compared with future pediatric study samples, particularly if the same test protocols are used.

Cognitive abilities change dramatically from the time children are entering preschool to when they reach young adulthood. These age-related changes do not typically exhibit linear or other simple polynomial relationships. The complex relationship between these maturational changes in cognitive abilities and corresponding changes in brain structures and neural networks has been examined in studies that employ multiple MRI modalities (Casey, Tottenham, Liston, & Durston, 2005; Jernigan, Baare, Stiles, & Madsen, 2011). After controlling for age, a variety of factors also contribute to differences in neurocognitive performance across individuals. Cognitive abilities across and within various developmental periods can vary between males and females. Sociodemographic factors, such as self-reported family income and race/ethnicity, help to explain a significant proportion of the variance in test performance (Mezzacappa, 2004; Noble, McCandliss, & Farah, 2007; Noble, Norman, & Farah, 2005; Waber, Carlson, Mann, Merola, & Moylan, 1984; Waber et al., 2007; Waber, Forbes, Almli, & Blood, 2012). Level of formal education is a significant predictor of test performance in adulthood (e.g., Heaton, Miller, Taylor, & Grant, 2004; Heaton, Taylor, & Manly, 2003) but age and education are almost totally confounded in children and adolescents. Parental level of education is a significant predictor of IQ and academic achievement in children (Breslau et al., 2001; Cirino et al., 2002) and therefore represents a more useful variable for investigation in pediatric studies. Failure to account for these factors in the interpretation of test scores can bias conclusions in clinical settings and threaten the validity of study results. As such, the influence of these various factors (sex, parental education, family income, and race/ethnicity) on age-related changes in NTCB was a critical focus of the current study. Specifically, we hypothesized that NTCB scores would show nonlinear effects of age, and that models that include this type of sociocultural information would perform significantly better than those that do not in the prediction of NTCB scores.

Estimating the effects of sociocultural factors within a highly diverse sample such as PING is challenging. The multi-site design produced a sample in which participants came from many different ethnic communities and many had mixed backgrounds. Because genotype information was available for the PING participants, we chose to use a set of genetically derived estimates of racial ancestry to estimate effects that could reflect differences in sociocultural background. We utilized a general additive model (GAM) regression methodology to provide flexible, data-driven estimates of the relationship between outcomes and age, rather than a strictly parametric approach such as polynomial regression. GAMs allow for potential nonlinear effects of independent variables on outcomes; the nature of the nonlinearity and degree of smoothness is data-determined for each independent variable (Hastie & Tibshirani, 1986; Wood, 2006).

Method

Participants

Participants were recruited through local postings and outreach activities conducted in the greater metropolitan areas of Baltimore, Boston, Honolulu, Los Angeles, New Haven, New York, Sacramento, and San Diego. The human research protections programs and institutional review boards at the 9 institutions participating in the PING project approved all experimental and consenting procedures. For individuals under 18 years of age, parental informed consent and child assent (for those 7 to 17 years of age) were obtained. All participants age 18 years and older gave their written informed consent.

Participants were excluded if there was a reported history of major developmental, psychiatric, or neurological disorders, brain injury, prematurity (i.e., born at less than 36 weeks gestational age), exposure to illicit drugs or alcohol prenatally for more than one trimester, history of head trauma with loss of consciousness for more than 30 minutes, or other medical conditions that could affect development. Individuals with contraindications for MRI studies (such as dental braces, metallic or electronic implants, claustrophobia, or pregnancy) were also excluded from participating. Individuals with identified or suspected learning disability or ADHD were not excluded since these syndromes are fairly common in pediatric populations.

To maintain comparability across analyses, only participants with valid data across all 7 NTCB tests (8 NTCB scores) were included in this study. Data from 118 participants (mean age = 6.33 years) were excluded because they failed to meet the minimum performance criteria (described below) on the Dimensional Change Card Sort Test (n=108), Flanker Inhibitory Control and Attention Test (n=9), and/or List Sorting Working Memory Test (n=13).

Information about socioeconomic status (SES) for each participant was based on the parent's indication of 'highest level of parental education' and 'family annual income' on the PING Study Demographics and Child Health History Questionnaire (participants age 18 to 20.9 were given a self-report version of this questionnaire). Highest level of parental education was categorized into 7 levels (from 'Less than seven years of school' to 'Professional'). The measure was defined as the highest level among those reported for either parent or guardian. Family annual income was categorized into 12 levels (from '< $5,000' to '$300,000 and above').

The final sample included 1020 participants between ages 3.0 and 20.9 years from the PING database as of July 5, 2012 who had complete data for the variables of interest. Sample recruitment was distributed across age and sex and was diverse in SES characteristics (see Tables 1 and 2).

Information about race and ethnicity was also collected on the PING Study Demographics and Child Health History Questionnaire. For those participants indicating a single racial category, 53.14% were White, 13.14% were African American/Black, 9.71% were Asian, 1.08% Native Hawaiian/Pacific Islander, and 0.88% American Indian/Alaskan Native. The remaining 22% indicated more than one racial category or "Other". Across this sample, 23% of the participants indicated that they were Hispanic/Latino.

To examine and control for the influences of race/ethnicity on test performance, genetic ancestry factors (GAFs) were calculated to estimate the proportion of European, African, American Indian, East Asia, Central Asia and Oceania ancestry for each participant, based on genotype analysis (methods detailed below).

NIH Toolbox Cognition Battery Measures

The validation study version of the NIH Toolbox Cognition Battery was utilized for this study and comprised 7 tests that measure 8 abilities within 6 major cognitive domains (Table 3). Details about the development of the test instruments and reliability and validity data for children ages 3 to 15 years are available (Weintraub, Bauer, et al., 2013; Weintraub, Dikmen, et al., 2013).

Prior to initiation of data collection the research coordinator from each PING testing site was trained in December 2009 by staff members from the Toolbox Project at Northwestern University. Testing at each PING site was conducted using a standard laptop computer and

touchscreen monitor. In order to maintain comparability of reaction times across trials during the Dimensional Change Card Sort and Flanker Inhibitory Control and Attention Tests, participants were instructed during the touchscreen tutorial to place their finger on a blue dot sticker on the table in front of them ('home base') prior to initiation of each trial. Test order was standardized across all participants following the instructions from the test developers: Dimensional Change Card Sort Test (DCCS), Flanker Inhibitory Control and Attention Test, Picture Sequence Memory Test, Pattern Comparison Processing Speed Test, Oral Reading Recognition Test, List Sorting Working Memory Test, and Picture Vocabulary Test. The scores for each measure were provided by the Toolbox Project staff at Northwestern University.

Dimensional Change Card Sort Test (DCCS)—The DCCS is the NTCB measure of cognitive flexibility or set shifting. The card sorting version of this test has been used to study the development of executive function in childhood (Beck, Schaefer, Pang, & Carlson, 2011; Zelazo, 2006). Participants were shown pictorial stimuli on the touchscreen monitor and instructed to match the central test stimuli with one of two lateralized target stimuli on the basis of either shape or color. Test trials consisted of a pre-switch block of 5 trials to be sorted by the last dimension used in the practice block, a post-switch block of 5 trials to be sorted by the other dimension, and a mixed block consisting of 50 trials, including 40 'dominant' and 10 'non-dominant' trials presented in a pseudorandom (fixed) order. Participants had to get 4 out of 5 trials correct in each of the pre- and post-switch blocks to proceed to the next level. The dominant dimension corresponded to that presented in the post-switch block. DCCS scoring was based on the pre- and post-switch blocks and the first 30 trials of the mixed block. A two-vector method was used that incorporated both accuracy and reaction time (RT) for participants who maintained a high level of accuracy (> 80% correct), and accuracy only for those who did not meet this criteria. Each vector score ranged from 0 to 5, for a maximum total score of 10.

Flanker Inhibitory Control and Attention Test—The NTCB version of the Eriksen flanker test (Eriksen & Eriksen, 1974) was adapted from the Attention Network Test (ANT; Rueda et al., 2004). Participants were required to indicate the left-right orientation of a centrally presented stimulus while inhibiting attention to the potentially incongruent surrounding stimuli (i.e., the flankers). The orientation of the flanking stimuli was congruent with the orientation of the central stimulus on some trials and incongruent on others. 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*). On congruent trials, there is no conflict and no obvious need for inhibitory control (that is, no need for more inhibitory control than is required simply to sustain attention on the task). Performance on these trials can thus be used as an index of *sustained attention*. The stimuli are divided into two blocks; fish (designed to be more engaging, as well as larger to make the task easier for younger children) and arrows (typical presentation for adults). The test consisted of of 25 fish trials, with 16 congruent and 9 incongruent trials presented in pseudorandom order. Participants who responded correctly on 5 or more of the 9 incongruent trials then proceeded to the arrows block. All children age 9 and above received both the fish and arrows blocks regardless of performance. The Flanker task yielded two scores; an inhibitory control score based on performance on both congruent and incongruent trials, and an attention score based on performance on congruent trials only. Both scores were derived using a two-vector procedure analogous to that used for the DCCS.

Picture Sequence Memory Test (PSMT)—In the PSMT, sequences of pictured objects and activities are presented on a computer screen. There is no inherent order within each sequence of thematically related pictures. The pictures are presented in a specific order that

the participant must remember and then reproduce by touching each of the pictures on the touchscreen and placing them in the correct order. Additional details are provided elsewhere (Bauer et al., 2013).

Three test trials were administered to each participant with the level of difficulty adjusted for different age ranges. A longer sequence was presented on trials 2 and 3 if perfect performance was obtained on the first test trial. Children ages 3 to 5 years were given 6 items (9 items for the longer sequence); ages 5 to 7: 9 items (12 items for the longer sequence); ages 7 to 9: 12 items (15 items for the longer sequence); and ages 9 and older: 15 items (18 items for the longer sequence). The participant's score on PSMT was derived from the cumulative number of adjacent pairs of pictures (i.e., two adjacent pictures placed in consecutive, ascending order) correctly recalled over 3 test trials. As a result, the maximum scores that children in each of these age groups could earn over 3 test trials were 21, 30, 39, and 48, respectively.

List Sorting Working Memory Test—The List Sorting Working Memory Test involves size order sequencing of familiar stimuli (Tulsky et al., 2013). Participants are presented with a series of illustrated pictures on the computer screen along with a recording of the name of the object. The test is divided into the One-List and Two-List conditions. In the One-List condition, participants are told to remember a series of objects (either food or animals) and repeat them in size order, smallest to largest. In the Two-List condition, participants are told to remember a series of objects (food and animals intermixed) and then respond by reporting the food in size order, followed by the animals in size order. Test items for both conditions were presented on the computer screen in a sequential manner. Each test item was a set of 2 strings (Item A & Item B); B was only administered if A was failed. Testing continued until two trials of the same series length were failed. The final List Sorting score consisted of combined total items correct on the One-List and Two-List conditions of the test (maximum = 28 points).

Picture Vocabulary Test—This measure of receptive vocabulary is administered in a computerized adaptive format. Preliminary item calibration was conducted online with 3,190 children ages 3 to 17 (Gershon et al., 2013). The participant is presented with an auditory recording of a word and four images on the computer screen; the task is to touch the image that most closely represents the meaning of the word. High-resolution color photos selected from the Getty Images library were used as stimuli. The Picture Vocabulary Test included 2 practice items (with feedback from the examiner) followed by 25 test items, the difficulty of which depends on the participant’s initial performance. Participant performance was converted to a Picture Vocabulary theta score (ranging from 4 to −4), based on item response theory (Linacre, 2005).

Oral Reading Recognition Test—In the NTCB version, a word or letter is presented on the computer screen and the participant is asked to read it aloud. Responses are recorded as correct or incorrect by the examiner, who views accepted pronunciations on a separate computer screen. For “pre-readers” and those with low literacy levels, letters and other multiple-choice “pre-reading” items are presented, to best assess the full range of ability across the age range, making the test as accessible as possible for young children. “Ceiling” rules were also implemented to minimize frustration, especially for pre-readers and early readers.

Each participant sees a series of letters and words presented one at a time on the computer screen. Items are presented in order of increasing difficulty. Four test forms were created from 280 items. Depending on age and prescreening performance, between 70 and 125 items were administered to each participant (the adaptive format version of this test was not yet

available). Children under the age of 8 were administered Form 1. Form assignment for older participants was based on their performance on the prescreening task, which consisted of 9 words and was discontinued after 3 consecutive errors. Once the form assignment was determined, the computer administered each item one by one, in an untimed fashion, until the last item in the assigned form was completed or the participant made 5 consecutive errors. The Oral Reading score is the total number of items correct.

Pattern Comparison Processing Speed Test—This test is designed to measure processing speed. Participants are required to discern whether two side-by-side pictures are the same by making a "yes" or "no" decision. Participants ages 3 to 7 were instructed to touch a smiley face on the touchscreen to correspond with a "yes" response and a frowning face to correspond with a "no" response; participants age 8 and older touched a "yes" or a "no" response button on the screen. Before testing, the younger participants were administered 2 sample items and 5 practice items; older participants were administered 6 practice items. After 90 seconds, the test trials were automatically discontinued. The Pattern Comparison score was the sum of correct responses within the 90 sec time limit.

Genetic Ancestry Assessment

We constructed a reference panel by collating genotype data collected for 2,513 individuals of known ancestry from 63 populations around the world using four publicly available sources (the Human Genome Diversity Project (HGDP; Cann et al., 2002), the Population Reference (POPRES; Nelson et al., 2008), the International HapMap 3 Consortium (HapMap3; Altshuler et al., 2010), and the University of Utah dataset (Xing et al., 2009)). The reference panel was created in a stepwise fashion in order to ensure that the included individuals not be admixed among the six continental populations, and that each continental population be represented by a reasonably large number of diverse individuals originating in the relevant continent. The assembled reference panel contained genotype information at 16,433 strand-unambiguous SNPs. These markers exhibited low LD (r-squared less than 0.1 was observed between 99% of marker pairs) and allele frequency higher than 1%.

To assess ancestry and admixture proportions in the PING participants, we used a supervised clustering approach implemented in the ADMIXTURE software (Alexander, Novembre, & Lange, 2009) and clustered participant data into six clusters corresponding to six major continental populations: African, Central Asian, East Asian, European, Native American, and Oceanic. These populations were defined by the individuals who comprised the previously described reference panel.

Estimating admixture proportions using a finite sample necessarily produces estimates that exhibit a certain level of error due to a sampling bias. However, this error can be estimated via bootstrapping. We developed a technique to reduce the noise associated with the admixture proportions by using the errors to refine the admixture estimates. In this approach, we first computed admixture estimates for all individuals using the entire set of reference individuals and determined the ancestry estimates' standard errors via bootstrapping. In the second step, we determined which individuals from the reference panel significantly contributed to estimation of each PING participants' ancestry based on 95% confidence intervals of the relevant admixture proportions. We then used this subset of individuals from the original reference panel in a subsequent supervised ADMIXTURE analysis to refine the initial admixture estimates. Final admixture estimates were calculated as the average of three separate runs with varying seeds as input for ADMIXTURE. The range of genetic ancestries observed in this sample is summarized in Table 4. As shown in Table 5, there is a high degree of agreement between genetically determined ancestry and self-reported ancestry.

Statistical Analyses

We used general additive models (GAMs) to analyze the relationship between age and other variables on the 8 NTCB scores. GAM is a regression methodology similar to the commonly used multiple linear regression, with the exception that specified independent variables or covariates are allowed to have smooth, nonlinear relationships with the dependent variable (Hastie & Tibshirani, 1986). The type of nonlinearity and degree of smoothness are not pre-specified but rather data determined. GAMs replace the assumption of linearity with the assumption of smoothness of the regression fits with age, where "smoothness" is defined as total variation in local curvature as a function of age (Hastie & Tibshirani, 1990). The degree of smoothing is empirically determined by minimizing an estimate of the prediction-squared error. The model is a procedure that iteratively applies a local regression scatterplot smoother to partial residuals controlling for other variables in the model (Hastie & Tibshirani, 1986). Compared to a quadratic fit, this approach is more robust to artificial effects (i.e., artificial peaks and no asymptotes) that have been observed in developmental data in other contexts (Fjell et al., 2010). GAMs can also incorporate traditional linear terms in addition to the smooth terms. In these analyses age was entered as a smoothly-varying independent variable and the covariates (sex, SES, GAFs) were entered as linear variables. For each NTCB score, a chi-square test was performed to determine whether the effect of age was non-linear or if a linear term was sufficient to describe the dependence of the NTCB score on age. Subsequently, we fitted three nested GAMs for each NTCB score, each subsequent model including all of the terms in the prior model. The base model consisted of a smooth of age, a linear term for sex, and a smooth interaction between sex and age. The second model included the two separate SES variables (i.e., parental education and household income) in addition to the base model; SES variables considered were highest education of the parents and annual family income, both added as linear terms. The third model added genetic ancestry factors (GAFs) as linear terms. For each pair of nested models we computed R-squared statistics and performed chi-square tests to determine whether the added terms represented a significant contribution over and above the terms already in the model.

Because the NTCB measures used in this study are raw scores, and they were administered to such a wide age range, it is expected that age will account for a large proportion of the variability in NTCB scores, making the contribution of SES and GAFs appear less meaningful than they might otherwise be when looking at children of the same age. Thus, we computed Cohen's f^2 effect size estimates (Cohen, 1988) to compare the base+SES +GAF model to the base model. Cohen's f^2 is an effect size used to estimate the proportion of explained (vs. unexplained) variation uniquely accounted for by a set of independent variables over and above that accounted for by all other variables in the model (Selya, Rose, Dierker, Hedeker, & Mermelstein, 2012). As such, it provides a more meaningful estimate of the contribution of these additional sociocultural factors when age is already accounted for. Cohen's f^2 is interpreted by convention in terms of small (.02), medium (.15), or large (.35) effects (Selya et al., 2012).

Results

Figure 1 shows the nonlinear age functions for each NTCB test score by sex. As expected, there were large and highly significant increases in the 8 NTCB scores with age. All models indicated a significant improvement when age was included as a smooth non-linear term in the model compared with models that included age as a linear term. The magnitude of increase in DCCS, Flanker, Attention, PSMT, and List Sorting scores with age was stronger for the younger children than the older children and adolescents (Figure 1, panels A to E). Scores on the Picture Vocabulary, Oral Reading Recognition, and Pattern Comparison

Processing Speed tests showed strong positive increases with age into adolescence (Figure 1, panels F to H).

When adjusted for age, sex explained a significant but small additional portion of individual variance in the DCCS (2.2%, $p = .00029$), List Sorting (2.1%, $p = .022$) and Pattern Comparison (2.4%, $p = .0053$) scores (Figure 1, panels A, E, and H). Males had significantly lower scores on the DCCS and Pattern Comparison tests than females, while males had significantly better performance than females on the List Sorting test. Performance across the other 5 NTCB scores showed no significant age x sex interactions and there were no significant sex differences in performance across the other 5 NTCB scores.

Table 6 shows R^2 statistics for the general additive models (GAMs) for each NTCB measure. The base model (age and sex) accounts for 53.8% (PSMT) to 73.4% (Oral Reading Recognition) of the variability in NTCB scores. The addition of SES to the base model accounts for an additional 1% to 2% of the variance in most measures, with the exception of Picture Vocabulary, for which it accounts for an additional 6.3% of the variance. The addition of genetic ancestry factors (GAFs) to the base+SES accounts for an additional .5% to 1% of the variance over and above SES for most NTCB scores, except for Vocabulary, for which it contributes an additional 2.3%. Table 7 shows the chi-square tests used to compare nested models. Chi-square statistics for the comparison of base and base+SES models indicate that the contribution of SES is significant at the $p < .0001$ level for all NTCB measures. For the comparison between base+SES and the base+SES+GAF models, the contribution of GAF is significant at the $p < .0001$ level for all NTCB measures except Flanker, DCCS, and Pattern Comparison (which were still significant at the $p < .001$ level).

The f^2 statistics are shown in Table 8 for comparison of the base+SES+GAF with the base model. In most cases the effect sizes associated with the addition of SES and GAF to the model were small. However, the effect sizes for the Oral Reading Recognition Test ($f^2 = .137$) and the Picture Vocabulary Test ($f^2 = .317$) were medium and large, respectively.

Discussion

These results demonstrate the age-related variation in performance on the NIH Toolbox Cognition Battery across 1020 typically developing children, adolescents, and young adults from ages 3 to 20 who were diverse in terms of SES and race/ethnicity. These results clearly demonstrate that performance on the NTCB does not change with age in a simple linear or polynomial manner. The use of general additive models appeared to have important advantages for estimating and describing the age-related changes in performance across this set of neuropsychological variables in a pediatric sample.

Because age accounted for such a large portion of the variability in NTCB scores, the relative contributions of SES and race/ethnicity (genetic ancestry factors (GAFs)) appear small (changes in R^2 ranging from 1% to 6% depending on the test and model). Clinicians and researchers are more accustomed to comparing performance among patients or study participants with that of other children of the same age and therefore looking at the contribution of these sociocultural factors in the context of this age variability may not be very useful. As such, we used the Cohen's f^2 effect size to estimate the importance of considering sociocultural information when predicting performance of the NTCB measures after age and sex have already been accounted for. When comparing the models in this way, the addition of SES+GAF added relatively little to the prediction of most of the NTCB scores. However, the addition of these variables made an important contribution to prediction of the Oral Reading Recognition and Picture Vocabulary Test scores, resulting in medium and large improvements in the prediction of these scores, respectively. These

results suggest that sociocultural factors may be particularly important when interpreting performance on language measures in children and adolescents. Researchers and clinicians frequently use measures of vocabulary as a proxy for general intelligence, and those using the NTCB may use the Picture Vocabulary Test as such a proxy in the future. Here we show that this measure is particularly sensitive to variables likely to index sociocultural factors; i.e., an estimated 32% of the variance in performance on the Picture Vocabulary Test not accounted for by age and sex was shared with the SES and GAF measures.

The overall results appear to be similar to those reported for a smaller sample of 208 typically developing children (120 3- to 6-year-olds and 88 8- to 15-year-olds) in the validation study (Weintraub, Bauer, et al., 2013). A revised version of the NIH Toolbox Cognition Battery is currently available at no or minimal cost to researchers (www.nihtoolbox.org) with normative data available from over 2500 participants for each year of age from 3 to 17, and for adults ages 18 to 85. In the future we plan to transform the NTCB scores from the entire PING dataset in order to provide researchers with results that are comparable with this revised version of the NTCB. These transformations are likely to be minimal in most cases. However, the test developers elected to use the keyboard rather than a touchscreen for input in this revised version, making transformation of scores difficult for the three speeded tasks (Dimensional Change Card Sort Test, Flanker Inhibitory Control and Attention Test, and Pattern Comparison Processing Speed Test). It is typically quite challenging to engage young children in reaction time tasks and obtain valid results (Akshoomoff, 2002), and there is continuing discussion in the neuropsychological community regarding the use of a touchscreen in testing. Although the standard method of administration in the new version of the NTCB will not use a touchscreen, examiners may still have the option of using a touchscreen if they are not concerned about standardizing scores using normative data that was collected without it. This may be an attractive option to some investigators, as observations by PING investigators suggest that the touchscreen may be significantly more engaging for younger children than the use of a mouse, keyboard, or button box. In this case, PING data may also be useful as alternative normative data for touchscreen administration.

The NTCB version of the DCCS appears to be significantly limited in its utility for measuring cognitive flexibility in children under age 7. Over 40% of the 3- to 6-year-olds in this sample were unable to execute the switching demands of this test, thereby failing to meet the performance criteria during the practice trials. A smaller percentage of younger children were unable to meet the practice criteria for the Flanker Inhibitory Control and Attention Test and the List Sorting Working Memory Test. These limitations were also reported in the validation study (Zelazo et al., 2013). In addition, the younger children had relatively poor performance during the test trials for these three tests. This resulted in scores on the DCCS, Flanker, and Attention measures for many of the youngest children that reflected only accuracy while the scores for the older participants were weighted for their speed of processing. The resulting bimodality and discontinuity in the age-distribution of the scores is clearly visible in Figure 1 (panels A, B, and C). It is unclear how these distributional anomalies affect the sensitivity of these measures in children under the age of 7. A substantial portion of this PING subsample also appeared to have ceiling level performance on the PSMT (Figure 1, panel D). These results indicate that longer sequences should have been available for older children and adolescents who achieved perfect performance on the lists that are presented. Test results are scored using item response theory in the revised version of the PSMT in the NTCB and may improve this limitation. Like the Picture Vocabulary test, the revised version of the Oral Reading Recognition test uses a computerized adaptive format and provides a theta score based on item response theory. This modification is likely to reduce administration time while providing an accurate estimate of the individual's oral reading abilities.

This study also has some limitations. The number of younger children included in this sample, particularly those between the ages of 3 and 5 years, was relatively small. This was primarily due to the fact that it was more challenging to recruit young children who were willing and able to cooperate with the MRI requirements of the PING study. This may have biased the generalizability of the results from the younger participants to some degree.

In summary, here we present the neuropsychological test data from the NTCB acquired within the Pediatric Imaging, Neurocognition and Genetics (PING) Study. This represents a unique dataset comprising behavioral measures from a sample of typically developing children spanning the age range from preschoolers to young adults. It is the largest dataset using the new NTCB in children and adolescents yet to be reported. The results highlight both the remarkable strengths and a few limitations of these new computerized assessment measures for use in neurodevelopmental research. The strengths include their sensitivity to neurodevelopmental effects and their efficiency. In addition to the expected age effects on performance, these results also indicate the particular impact of sociocultural variables on performance for two of the measures. Limitations are observed in the form of some ceiling effects in older children on some tasks and floor effects, particularly on executive function tests, in the youngest participants.

In the PING Data Resource, these neuropsychological data are accompanied by multimodality imaging measures and high-density SNP genotyping information. PING data will be made available to the entire research community for data mining and data exploration in the near future. This project is among the first large coordinated projects to address the important aim shared by both the National Institutes of Health and the research community generally to create larger, more powerful data repositories through data sharing. Such initiatives are critical to our success in increasing the scientific yield of research dollars spent to address shared aims. It is likely that answers to many of the most important questions posed in neurodevelopmental research can best be obtained from complex, multidimensional models feasible only in the context of very large corpora of data. PING consortium investigators fully endorse and are observing first hand the value of such data sharing initiatives.

Acknowledgments

The authors gratefully thank the children, adolescents, young adults, and parents who participated in this study. We thank Tyler V. Smith for his role in implementing the PING database. In the PING Genomics Core, Burcu Darst coordinated samples and Rebecca Tisch and Nikki Villarasa were responsible for DNA isolation and genotyping.We thank our research staff who were responsible for subject recruitment and testing, particularly Da Yea Jang, Shereen Cohen, Bella Eyngorina, Erika Ruberry, Alisa Powers, Natasha Mehta, Alexandra Pritchett, Shannon Kogachi, Kristin Lee, Erin Fukaya, Antonette Hernandez, Joan Bosson-Heenan, Holly Manigan, Lauren Brodsky, Nitzah Gebhard, Suzanne Houston, Lindsay Soderberg, Diane Lanham, Daina Crafa, Hilda Rizzo-Busack, Lauren Yakutis, and Michael Hill. Manning Richardsonassisted WKT with statistical analyses. We are grateful to the NIH Toolbox staff at Northwestern University (particularly Katy Wortman, Katy Browne, Ed Bedjeti, and Dr. Cindy Nowinski), as well as Drs. Cheryl Boyce and Jim Bjork at the National Institute on Drug Abuse and Dr. Molly Wagster at the National Institute of Aging for their support of this project.

Data used in preparation of this study were obtained from the PING database. As such, the investigators within PING contributed to the design and implementation of PING and/or provided data but did not necessarily participate in the analysis or writing of this report. A complete listing of PING investigators can be found at http://ping.chd.ucsd.edu. Data collection and sharing for this project was funded by the National Institute on Drug Abuse and the Eunice Kennedy Shriver National Institute of Child Health and Human Development grant RC2DA029475.

References

Akshoomoff NA. Selective attention and active engagement in young children. Developmental Neuropsychology. 2002; 22:625–642. [PubMed: 12661973]

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19(9):1655–1664. [PubMed: 19648217]

Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, McEwen JE. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467(7311):52–58. [PubMed: 20811451]

Bakken TE, Roddey JC, Djurovic S, Akshoomoff N, Amaral DG, Bloss CS, Carlson H. Association of common genetic variants in GPCPD1 with scaling of visual cortical surface area in humans. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109(10): 3985–3990. [PubMed: 22343285]

Bauer PJ, Dikmen SS, Heaton RK, Mungas D, Slotkin J, Beaumont JL. Iii. Nih toolbox cognition battery (cb): measuring episodic memory. Monogr Soc Res Child Dev. 2013; 78(4):34–48.

Beck DM, Schaefer C, Pang K, Carlson SM. Executive Function in Preschool Children: Test-Retest Reliability. J Cogn Dev. 2011; 12(2):169–193. [PubMed: 21643523]

Breslau N, Chilcoat HD, Susser ES, Matte T, Liang KY, Peterson EL. Stability and change in children's intelligence quotient scores: a comparison of two socioeconomically disparate communities. American Journal of Epidemiology. 2001; 154(8):711–717. [PubMed: 11590083]

Brown TT, Kuperman JM, Chung Y, Erhart M, McCabe C, Hagler DJ Jr, Dale AM. Neuroanatomical assessment of biological maturity. Current Biology. 2012; 22(18):1693–1698. [PubMed: 22902750]

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Cavalli-Sforza LL. A human genome diversity cell line panel. Science. 2002; 296(5566):261–262. [PubMed: 11954565]

Casey BJ, Tottenham N, Liston C, Durston S. Imaging the developing brain: what have we learned about cognitive development? Trends Cogn Sci. 2005; 9(3):104–110. [PubMed: 15737818]

Cirino PT, Chin CE, Sevcik RA, Wolf M, Lovett M, Morris RD. Measuring socioeconomic status: reliability and preliminary validity for different approaches. Assessment. 2002; 9(2):145–155. [PubMed: 12066829]

Cohen, J. Statistical power analysis for the behavioral sciences. 2nd. Hillsdale, NJ: 1988.

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.

Fjell, AM.; Walhovd, KB.; Brown, TT.; Kuperman, JM.; Chung, Y.; Hagler, DJ., Jr; Dale, AM. Multimodal imaging of the self-regulating developing brain; Proceedings of the National Academy of Sciences of the United States of America; 2012.

Fjell AM, Walhovd KB, Westlye LT, Østby Y, Tamnes CK, Jernigan TL, Gamst A, Dale AM. When does brain aging accelerate? Dangers of quadratic fits in cross-sectional studies. Neuroimage. 2010; 50:1376–1383. [PubMed: 20109562]. [PubMed: 20109562]

Gershon RC, Cella D, Fox NA, Havlik RJ, Hendrie HC, Wagster MV. Assessment of neurological and behavioural function: the NIH Toolbox. Lancet Neurol. 2010; 9(2):138–139. [PubMed: 20129161]

Gershon RC, Slotkin J, Manly JJ, Blitz DL, Beaumont JL, Schnipke D, Weintraub S. iv. NIH Toolbox Cognition Battery (cb): measuring language (vocabulary comprehension and reading decoding). Monogr Soc Res Child Dev. 2013; 78(4):49–69.

Hastie T, Tibshirani R. Generalized additive models. Statistical Science. 1986; 1(3):297–318.

Hastie, T.; Tibshirani, R. Generalized Additive Models. London: Chapman and Hall; 1990.

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.; 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.

Jernigan TL, Baare WF, Stiles J, Madsen KS. Postnatal brain development: structural imaging of dynamic neurodevelopmental processes. Progress in Brain Research. 2011; 189:77–92. [PubMed: 21489384]

Linacre, JM. A User's Guide to WINSTEPS/MINISTEP: Rasch-Model Computer Programs. Chicago, IL: Winsteps; 2005.

Mezzacappa E. Alerting, orienting, and executive attention: developmental properties and sociodemographic correlates in an epidemiological sample of young, urban children. Child Development. 2004; 75(5):1373–1386. [PubMed: 15369520]

Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Lai EH. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. American Journal of Human Genetics. 2008; 83(3):347–358. [PubMed: 18760391]

Noble KG, McCandliss BD, Farah MJ. Socioeconomic gradients predict individual differences in neurocognitive abilities. Dev Sci. 2007; 10(4):464–480. [PubMed: 17552936]

Noble KG, Norman MF, Farah MJ. Neurocognitive correlates of socioeconomic status in kindergarten children. Dev Sci. 2005; 8(1):74–87. [PubMed: 15647068]

Rueda MR, Fan J, McCandliss BD, Halparin JD, Gruber DB, Lercari LP, Posner MI. Development of attentional networks in childhood. Neuropsychologia. 2004; 42(8):1029–1040. [PubMed: 15093142]

Selya AS, Rose JS, Dierker LC, Hedeker D, Mermelstein RJ. A practical guide to calculating Cohen's f(2), a measure of local effect size, from PROC MIXED. Front Psychol. 2012; 3:111. [PubMed: 22529829]

Tulsky DS, Carlozzi NE, Chevalier N, Espy KA, Beaumont JL, Mungas D. v. NIH Toolbox Cognition Battery (cb): measuring working memory. Monogr Soc Res Child Dev. 2013; 78(4):70–87.

Waber DP, Carlson D, Mann M, Merola J, Moylan P. SES-related aspects of neuropsychological performance. Child Development. 1984; 55(5):1878–1886. [PubMed: 6510058]

Waber DP, De Moor C, Forbes PW, Almli CR, Botteron KN, Leonard G, Rumsey J. The NIH MRI study of normal brain development: performance of a population based sample of healthy children aged 6 to 18 years on a neuropsychological battery. Journal of the International Neuropsychological Society. 2007; 13(5):729–746. [PubMed: 17511896]

Waber DP, Forbes PW, Almli CR, Blood EA. Four-year longitudinal performance of a population-based sample of healthy children on a neuropsychological battery: the NIH MRI study of normal brain development. Journal of the International Neuropsychological Society. 2012; 18(2):179–190. [PubMed: 22364826]

Walhovd KB, Fjell AM, Brown TT, Kuperman JM, Chung Y, Hagler DJ Jr, Dale AM. Long-term influence of normal variation in neonatal characteristics on human brain development. Proceedings of the National Academy of Sciences of the United States of America. 2012

Weintraub S, Bauer PJ, Zelazo PD, Wallner-Allen K, Dikmen SS, Heaton RK, Gershon RC. i. NIH Toolbox Cognition Battery (cb): introduction and pediatric data. Monogr Soc Res Child Dev. 2013; 78(4):1–15. [PubMed: 23952199]

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):S54–64. [PubMed: 23479546]

Wood, SN. Generalized additive models : an introduction with R. Boca Raton, FL: Chapman & Hall/CRC; 2006.

Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Jorde LB. Mobile elements create structural variation: analysis of a complete human genome. Genome Research. 2009; 19(9):1516–1526. [PubMed: 19439515]

Zelazo PD. The Dimensional Change Card Sort (DCCS): a method of assessing executive function in children. Nat Protoc. 2006; 1(1):297–301. [PubMed: 17406248]

Zelazo PD, Anderson JE, Richler J, Wallner-Allen K, Beaumont JL, Weintraub S. ii. NIH Toolbox Cognition Battery (cb): measuring executive function and attention. Monogr Soc Res Child Dev. 2013; 78(4):16–33.

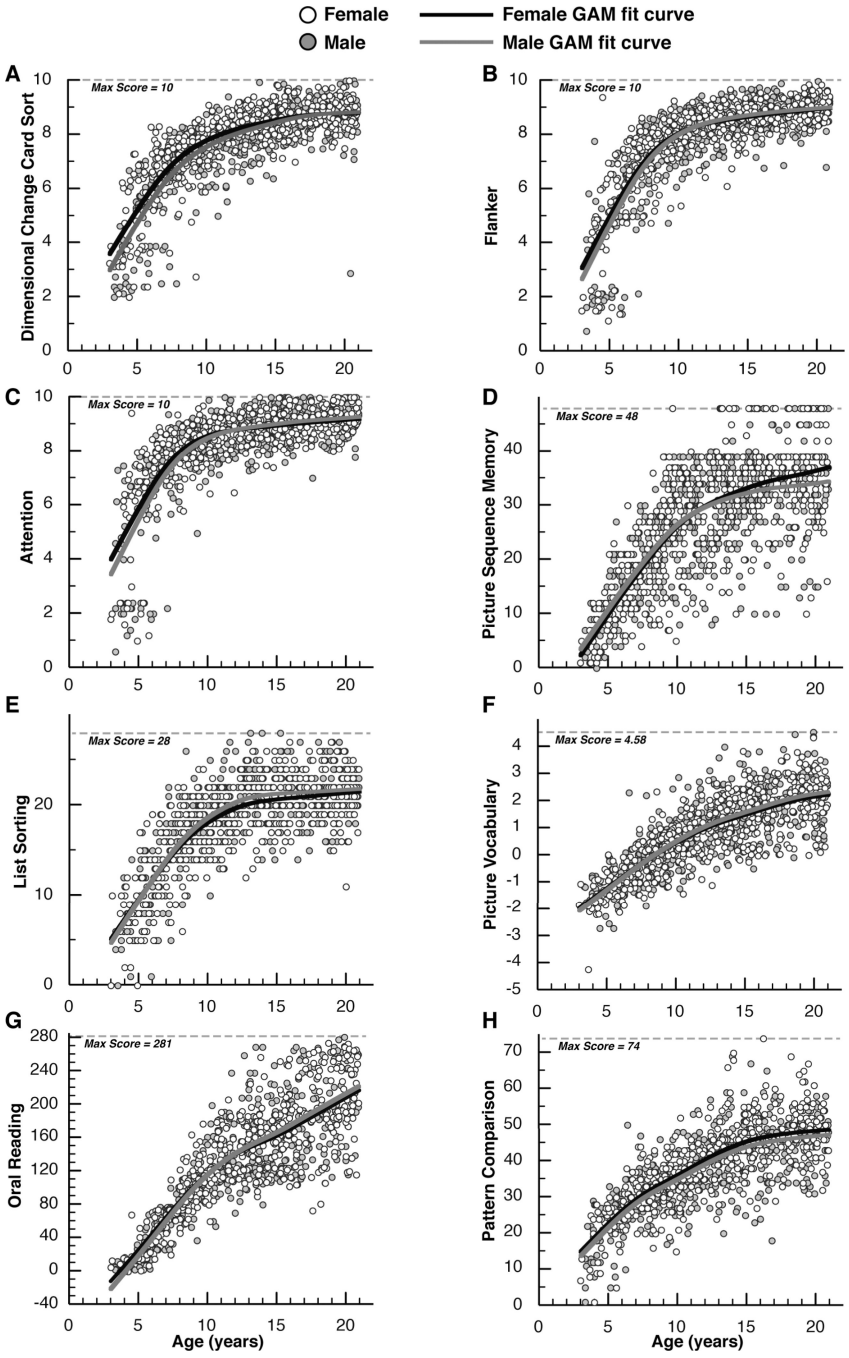

Figure 1.

NIH Toolbox Cognition Battery scores for individual subjects plotted against age. Lines show fit from GAMs by sex, controlling for SES and GAF by setting these values to their sample means.

Table 1

Sample age and sex characteristics.

Age Group	Mean Age (SD)	Males	Females	Total
3.0–4.9 years	4.17 (0.47)	24	24	48
5.0–6.9 years	5.98 (0.56)	50	50	100
7.0–8.9 years	7.96 (0.56)	68	61	129
9.0–10.9 years	9.94 (0.58)	73	71	144
11.0–13.9 years	12.55 (0.92)	114	82	196
14.0–16.9 years	15.48 (0.83)	106	87	193
17.0–19.9 years	18.61 (0.84)	71	73	144
20.0–20.9 years	20.44 (0.32)	28	38	66
Total:		534	486	1020

Table 2

Socioeconomic status characteristics of the study sample.

Highest Level of Parental Education	Portion of Participants
< 7 years of school	.6%
7 to 9 years of school	.6%
10 to 11 years of school	2.1%
High school graduate	11.3%
1 to 3 years college	23.9%
College graduate	27.7%
Professional (MA, MS, MD, PhD, etc.)	33.8%

Family Annual Income	Portion of Participants
< $5,000	3.5%
$5,000-$9,999	3.2%
$10,000-$19,999	6.0%
$20,000-$29,999	5.5%
$30,000-$39,999	7.0%
$40,000-$49,999	6.0%
$50,000-$99,999	29.9%
$100,000-$149,999	19.0%
$150,000-$199,999	9.4%
$200,000-$249,999	4.1%
$250,000-$299,999	2.2%
$300,000 and above	4.2%

Table 3

NIH Toolbox Cognition Battery Measures

Subdomain	Ability	Toolbox Measure
Executive Function	Cognitive Flexibility	Dimensional Change Card Sort Test
Executive Function	Inhibitory Control	Flanker Inhibitory Control and Attention Test
Attention	Visual Attention	Flanker Inhibitory Control and Attention Test
Episodic Memory	Episodic Memory	Picture Sequence Memory Test
Processing Speed	Processing Speed	Pattern Comparison Processing Speed Test
Language	Oral Reading Skill	Oral Reading Recognition Test
Working Memory	Working Memory	List Sorting Working Memory Test
Language	Vocabulary Knowledge	Picture Vocabulary Test

Table 4

Genetic ancestry characteristics for the PING sample. Participants were classified with an ancestral continental population if there was ≥ 80% genomic similarity.

Ancestral Continental Population	Number in PING Sample	% of Sample
European	486	47.6
African	86	8.4
East Asian	82	8.0
Central Asian	17	1.7
Native American	1	<1
Oceanic	0	0
Admixed [1]	348	34.1

[1] Genomic similarity to more than one continental population

Table 5

Spearman correlations between genetic ancestry factor (proportions) and self-reported ancestry (percent derived). Positive statistically significant correlations are shown in bold.

Genetic Ancestry Factor	Self-Reported Ancestry				
	White	**African American**	**Asian**	**Pacific Islander**	**American Indian**
European	**.846 p<.0005**	−.460 p<.0005	−.535 p<.0005	−.220 p<.0005	−.062 p=.047
African	−.516 p<.0005	**.757 p<.0005**	−.218 p<.0005	−.022 p=.489	.061 p=.051
Central Asian	−0.37 p=232	−.053 p=.091	**.148 p<.0005**	−.023 p=.471	−.044 p=.160
East Asian	−.464 p<.0005	−.167 p<.0005	**.780 p<.0005**	**.431 p<.0005**	.042 p=.176
Oceanic	−.285 p<.0005	−.069 p=.027	**.372 p<.0005**	**.710 p<.0005**	.034 p=.282
Amerindian	−.050 p=.107	−.102 p=.001	−.169 p<.0005	−.083 p=.008	**.187 p<.0005**

Table 6

Results from the GAMs (*** $p<.001$) to estimate the relationship between each of the 8 NIH Toolbox Cognition Battery scores and the base model (a smooth of age, a linear term for sex, and a smooth of the interaction between sex and age), the second model (base model + SES added as a linear term), and the third model (base model + SES and GAF added as linear terms).

	R^2 Base Model	R^2 Second Model	R^2 Third Model
DCCS Score	0.665***	0.680***	0.685***
Flanker Score	0.680***	0.690***	0.695***
Attention Score	0.582***	0.593***	0.601***
PSMT Score	0.538***	0.554***	0.570***
Pattern Comparison Score	0.572***	0.582***	0.588***
Oral Reading Score	0.734***	0.756***	0.766***
List Sorting Score	0.621***	0.641***	0.648***
Picture Vocabulary Score	0.643***	0.706***	0.729***

Table 7

Chi-square values for each pair of nested models demonstrating the significance of the addition of SES and GAF to the terms already in the model

	SES (*df*=2)	**GAF (*df*=5)**
DCCS Score	54.883***	18.364**
Flanker Score	39.535***	16.714**
Attention Score	31.557***	24.962***
PSMT Score	35.925***	38.664***
Pattern Comparison Score	23.492***	15.989**
Oral Reading Score	93.793***	44.974***
List Sorting Score	59.442***	23.256***
Picture Vocabulary Score	217.917***	88.701***

*** $p<.001$

** $p<.01$.

Table 8

Effect sizes (Cohen's f^2) for each NIH Toolbox Cognition Battery scores associated with the addition of SES+GAF to the base model.

DCCS	Flanker	Attention	PSMT	Pattern Comparison	Oral Reading	List Sorting	Picture Vocabulary
.064	.049	.048	.074	.039	.137	.077	.317