Developmental Cognitive Neuroscience 24 (2017) 72–83

Contents lists available at ScienceDirect

Developmental Cognitive Neuroscience

journal homepage: http://www.elsevier.com/locate/dcn

Effects of prior testing lasting a full year in NCANDA adolescents: Contributions from age, sex, socioeconomic status, ethnicity, site, family history of alcohol or drug abuse, and baseline performance☆,☆☆

Edith V. Sullivan[a,*], Ty Brumback[b], Susan F. Tapert[b,c], Devin Prouty[d], Rosemary Fama[a,d], Wesley K. Thompson[b], Sandra A. Brown[b], Kevin Cummins[b], Ian M. Colrain[d], Fiona C. Baker[d], Duncan B. Clark[e], Tammy Chung[e], Michael D. De Bellis[f], Stephen R. Hooper[g], Bonnie J. Nagel[h], B. Nolan Nichols[a,d], Weiwei Chu[d], Dongjin Kwon[d], Kilian M. Pohl[a,d], Adolf Pfefferbaum[a,d]

[a] Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, United States
[b] Department of Psychiatry, University of California, San Diego, La Jolla, CA, United States
[c] Veterans Affairs San Diego Healthcare System, La Jolla, CA, United States
[d] Center for Health Sciences, SRI International, Menlo Park, CA, United States
[e] Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, United States
[f] Healthy Childhood Brain Development Research Program, Department of Psychiatry & Behavioral Sciences, Duke University School of Medicine, Durham, NC, United States
[g] Department of Allied Health Sciences, School of Medicine, University of North Carolina, Chapel Hill, NC, United States
[h] Departments of Psychiatry and Behavioral Neuroscience, Oregon Health & Sciences University, Portland, OR, United States

ARTICLE INFO

Article history:
Received 6 September 2016
Received in revised form
28 November 2016
Accepted 9 January 2017
Available online 24 January 2017

Keywords:
Cognitive development
Motor development
Longitudinal
Alcohol
Adolescence
Practice effects

ABSTRACT

Longitudinal study provides a robust method for tracking developmental trajectories. Yet inherent problems of retesting pose challenges in distinguishing biological developmental change from prior testing experience. We examined factors potentially influencing change scores on 16 neuropsychological test composites over 1 year in 568 adolescents in the National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA) project. The twice-minus-once-tested method revealed that performance gain was mainly attributable to testing experience (practice) with little contribution from predicted developmental effects. Group mean practice slopes for 13 composites indicated that 60% to ~100% variance was attributable to test experience; General Ability accuracy showed the least practice effect (29%). Lower baseline performance, especially in younger participants, was a strong predictor of greater gain. Contributions from age, sex, ethnicity, examination site, socioeconomic status, or family history of alcohol/substance abuse were nil to small, even where statistically significant. Recognizing that a substantial proportion of change in longitudinal testing, even over 1-year, is attributable to testing experience indicates caution against assuming that performance gain observed during periods of maturation necessarily reflects development. Estimates of testing experience, a form of learning, may be a relevant metric for detecting interim influences, such as alcohol use or traumatic episodes, on behavior.

1. Introduction

1.1. Background

Advancement in ability to engage in complexity of thought and analysis characterizes maturation throughout adolescence. Despite the many test batteries available to measure such functional advancement, metrics to quantify developmental changes studied cross-sectionally or longitudinally are necessarily contaminated by numerous factors, including socioeconomic status (SES), geographically-linked educational opportunities, and ethnic-

☆ EVS received additional support from the Moldow Women's Hope and Healing Fund.
☆☆ This work was supported by the U.S. National Institute on Alcohol Abuse and Alcoholism with co-funding from the National Institute on Drug Abuse, the National Institute of Mental Health, and the National Institute of Child Health and Human Development [NCANDA grant numbers: AA021697 (AP+KMP), AA021695 (SAB+SFT), AA021692 (SFT+SAB), AA021696 (IMC+FCB), AA021681 (MDDB), AA021690 (DBC), AA021691 (BN); K05 AA017168 (EVS)].
* Corresponding author at: Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine (MC5723), 401 Quarry Road, Stanford, CA 94305-5723, United States.
E-mail address: edie@stanford.edu (E.V. Sullivan).

https://doi.org/10.1016/j.dcn.2017.01.003

related cultural differences, that are often secondary to primary questions about the differential rates of development for constellations of functions. An additional confounding factor in longitudinal study is prior experience with test material and procedures, commonly considered practice effects, which, if unaccounted, can be misinterpreted as developmental change.

Over the past several decades, a number of studies have been initiated to investigate the rate and pattern of different component functional processes and factors that influence their development. Among the recent large scale longitudinal, child and adolescent studies are the NIH MRI Brain Development Cooperative Group Developmental Study (Giedd et al., 2014; Waber et al., 2012), the Pediatric (Longitudinal) Imaging, Neurocognition, and Genetics Study (P[L]ING) (Akshoomoff et al., 2014; Jernigan et al., 2016), the IMAGEN study (Schumann et al., 2010), and the Philadelphia Neurodevelopmental Cohort (Gur et al., 2012; Satterthwaite et al., 2016). These studies have identified significant contributions to performance age-related differences or change from SES or site (Akshoomoff et al., 2014). Typically, the sexes do not differ in cognitive developmental trajectories, even where they differ in performance level (Cromer et al., 2015; Gur et al., 2012; Waber et al., 2012). Varying amounts of practice have been reported and depend on the function tested (Ibrahim et al., 2015; Waber et al., 2012) and test interval and can show effects of prior testing experience, even over a 2-year interval (Waber et al., 2012). One exceptionally large study used the Cogstate Brief Battery to assess four processes (psychomotor function, attention, working memory, and visual learning) in 38,778 adolescents cross-sectionally and 5788 adolescents (10–18 years old) longitudinally over a 1-year interval (Cromer et al., 2015). Little change was detected over the year, boys and girls performed similarly, and rate of responding slowed with older age; a practice effect was noted on only one learning test. Likewise, the Philadelphia study identified few prior-testing effects over 5 years and less variability in healthy controls and unaffected family members than in retested patients with schizophrenia; all were adults (Gur et al., 2010; Roalf et al., 2013).

"Practice effects" refer to prior experience with testing and not necessarily with specific test materials *per se*. Improvement over time is strongly related to initial performance level. For example, a longitudinal controlled study of children and adolescents with hemophilia used age-appropriate IQ batteries (Wechsler Intelligence Scale for Children-Revised and Wechsler Adult Intelligence Scale-Revised) over 4 years of annual testing (Sirois et al., 2002). Performance scaled scores improved over 2 years, whereas Verbal scaled scores declined slightly but insignificantly, regardless of whether controls were given the same test or aged into the adult version. Compared with the modest practice effect, baseline performance was the strongest predictor of later performance. A meta-analysis of about 1600 effect sizes of change scores and practice effects identified use of different test forms, age, diagnosis, and test-retest interval as contributing factors to change (Calamia et al., 2012). The mainstay of these studies was based on change scores in adulthood, disease, or senescence.

Among studies identifying clear differential contributions to change from childhood to early adolescent development relative to practice is one by Anderson et al. (Anderson et al., 2001). The report provided a re-analysis of a longitudinal study by Nettelbeck and Wilson (Nettelbeck and Wilson, 1985), who reported substantial maturational effects on an inspection time task with little consideration of potential practice effects. The re-analysis examined differences in performance over the four test sessions: baseline and retest conducted immediately, 1 year, and 2 years after baseline. Although little gain, averaging 1.18%, was measurable between baseline and the immediate retest, improvement from baseline to 1 year averaged nearly 30%; little further gain accrued at year 2. The calculation to quantify practice vs. development in this study was, in principle, the twice-minus-once-tested method (Salthouse, 2015).

1.2. Study aims

Our current study examined multiple factors influencing change scores on neuropsychological testing (Sullivan et al., 2016) over 1 year in 568 adolescents in the National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA) project (Brown et al., 2015) using general additive modeling. In particular, we focused on two study aims. The first was to distinguish and quantify developmental and prior experience (practice) effects on cognitive and motor measures assessing seven functional domains, represented by 14 composite scores, tested over a 1-year interval. Accordingly, we examined performance by NCANDA participants who had baseline and 1 year follow-up neuropsychological testing and remained in the no-to-low drinking and drug consumption group; that is, these adolescents met age- and sex-specific criteria for alcohol and drug consumption at study entry (Brown et al., 2015) and maintained that status at the 1-year examination. The neuropsychological battery used the same methods and test forms at both examinations, and the same composite scores were derived for both test sessions. The mainstay of the tests was drawn from the University of Pennsylvania Web Computerized Neuropsychological Battery (WebCNP) (Gur et al., 2012), thereby affording measures of accuracy and speed; traditional tests were also included in the battery (Sullivan et al., 2016). To estimate effects of prior testing experience, we used a version of the twice-minus-once-tested method by comparing longitudinally-determined change over the 1-year test interval and expected change estimated from cross-sectionally-determined age-dependent baseline averages (see Statistical Analysis for the formula). A second study aim was to assess the contributions of age, baseline test performance, sex, SES, ethnicity, family history of alcohol or substance use, and examination site to performance change.

2. Material and methods

2.1. Participants

The participants in this analysis included 568 adolescents (age 12–21 at entry) with baseline and 1-year followup data and were drawn from the 831 adolescents tested at baseline (Sullivan et al., 2016). The reduction in sample size from baseline to the 1-year follow-up reflects attrition and change in drinking status over the test interval. At baseline, the participants were recruited across five NCANDA sites: University of California at San Diego, SRI International, Duke University Medical Center, University of Pittsburgh Medical Center, and Oregon Health & Science University (Brown et al., 2015). Of the 831 participants, 692 met criteria for no/low alcohol or drug exposure at baseline, and 568 with test scores remained in the no/low group at the one-year follow-up.

2.1.1. Informed consent

All participants underwent informed consent processes at both visits with a research associate trained in human subject research protocols. Adult participants or the parents of minor participants provided written informed consent before participation in the study. Minor participants provided assent before participation. The Institutional Review Boards of each site approved this study, and each site followed this procedure to obtain voluntary informed consent or assent, depending on the age of the participant.

Table 1
NCANDA demographics for 1-year followup.

Age (years)	mean=	15.4
	SD=	2.35
	N=	568
Male	mean=	15.4
	SD=	2.31
	N=	289
Female	mean=	15.5
	SD=	2.39
	N=	279
Socioeconomic status[a]	mean=	16.7
	SD=	2.47
Family History of Alcoholism negative, positive = 489, 79		
Self-declared Ethnicity		
Caucasian	N=	413
African-American	N=	80
Asian	N=	64
Other	N=	11
Site		
UPitt	N=	78
SRI	N=	106
Duke	N=	116
OHSU	N=	114
UCSD	N=	154

[a] Highest education of a parent.

2.1.2. Subject demographics

As described in our baseline analysis (Sullivan et al., 2016), participants were characterized by age, sex, self-identified ethnicity, socioeconomic status (SES) determined as the highest level of education achieved by either parent (Akshoomoff et al., 2014), and family history of alcohol or drug use disorder. In light of the substantial differences in salaries and incomes across the five geographically-distributed data collection sites, we expressed SES with reference to parental education level, which is less subject than family income to geographical differences in the U.S. Most subjects reported a single self-identified ethnicity (Caucasian, African-American, Asian, Pacific Islander, and Native American) with some reporting mixed heritage. There were adequate numbers of the first three types to assign categorical ethnicity, with dual-heritage identifications assigned to the minority ethnicity group (e.g., Asian-Caucasian was categorized as Asian) (Table 1).

2.1.3. Alcohol history and testing

Participants completed the Customary Drinking and Drug use Record (CDDR, Brown et al.,1998) to characterize past and current alcohol and substance use. All participants also submitted samples to a 12-panel urine toxicology screen for amphetamine, methamphetamine, cocaine, phencyclidine, benzodiazepines, barbiturates, opiates, oxycodone, propoxyphene, methadone, tricyclic antidepressants, marijuana and a breathalyzer for alcohol to confirm absence of evidence for recent use of drugs of abuse. Positive screens other than marijuana were sent for GC/MS confirmation, and if confirmed, participants were excluded from testing.

2.2. Neuropsychological tests and procedures

Assessment was the same across all sites and used a combination of computerized tests [WebCNP (https://webcnp.med.upenn.edu/) (Gur et al., 2012; Gur et al., 2010)] and traditional neuropsychological tests (Sullivan et al., 2016). Testing was conducted by research assistants trained with annual reliability evaluations to criterion and calibrated annually by a centrally-trained psychometrician using procedures established by the NCANDA Data Analysis Component. The tests were administered in the same order across all sites and were generally completed in approximately 3 h. Test results were uploaded to the software platform, Scalable Informatics for Biomedical Imaging Studies (Nichols and Pohl, 2015; Rohlfing et al., 2014; SIBIS; https://github.com/sibis-platform) at SRI International. The longitudinal data used in this manuscript were organized via a formal, locked data release (NCANDA_RELEASE_00001_V01). Additional information about SIBIS, the data management system used by NCANDA, has been published elsewhere (Nichols and Pohl, 2015; Rohlfing et al., 2014).

The WebCNP has established construct validity and reliability and was standardized on upwards of 10,000 participants (depending on the measure) with a broad, age range (8–90 years old) (Gur et al., 2010). Descriptions of the 15 WebCNP tests used were provided in our earlier report (Sullivan et al., 2016) (Supplemental Table 3), with most tests having both accuracy and speed (response time) measures. The descriptions were modified from the WebCNP support manual. A subset of measures from these tests was used to create theoretically-driven composite Z-scores for 8 accuracy measures (General Ability, Abstraction, Attention, Emotion, Episodic Memory, Working Memory, Balance, and Total) and 8 speed measures (General Ability, Abstraction, Attention, Emotion, Episodic Memory, Working Memory, Motor, and Total). The individual tests and computed composites are summarized next and fully described elsewhere [(Sullivan et al., 2016), where Table 2 lists the cognitive and motor domains and specific processes assessed, with associated brain regions reported to support each process; Supplemental Table 2 lists the composite domains, test measures and variable names entered into each composite domain, and scoring procedure for each measure].

Composite score construction followed three steps (Gur et al., 2012; Sullivan et al., 1994). First, each measure was standardized on baseline scores achieved by all male and female adolescents who met NCANDA entry criteria (maximum N = 692) and expressed as a Z-score (mean = 0 ± 1SD). This transformation function was applied to all subjects at all times. Not all participants had scores for all measures, typically due to computer failure, participant's refusal to perform a test, or lack of testing time. Next, all scores for which a low score signified good performance were transformed by multiplying scores by −1 so that high scores for all measures were in the direction of good performance. Finally, the mean Z-score of all individual measures that comprised a composite was calculated; if a subject was missing a score, that composite score was not calculated for that subject.

2.3. Neuropsychological test composite composition

The test battery comprising numerous performance measures assessed 7 functional domains, enabling construction of hypothesis-driven composite accuracy scores. Of the 8 accuracy composites (based on the 7 functional domains and a total composite score), only the Balance composite score did not have a complementary speed or response score. Of the 8 speed composites, only the Motor Speed composite did not have a complementary accuracy score.

2.3.1. Abstraction

Conditional Exclusion measures abstraction and mental flexibility. There are three principles for choosing an object: line thickness, shape, and size. The target principle changes as the participant achieves 10 consecutive correct responses for each principle. There is only one principle in effect for any trial, but a response may match more than one principle. The participant is not told what the targeted principle is and must derive the correct principle through feedback. The participant has 48 trials to make 10 consecutive cor-

Table 2
GAM of the slopes for each composite.

Composite	N	Age at Baseline		Baseline Performance		SES[a]		Ethnicity		Sex		Site		Family History		Group mean practice slope = % group mean observed slope
		variance	p	variance	p	variance	p	variance	p	variance	p	variance	p	variance	p	
Accuracy Composites																
General Ability	545	0.0030	0.0918	**0.0513**	**0.0000**	**0.0183**	**0.0027**	0.0109	0.1046	0.0023	0.2549	0.0119	0.1335	0.0013	0.3909	29.17
Abstraction	542	**0.0263**	**0.0000**	**0.2534**	**0.0000**	**0.0185**	**0.0001**	**0.0219**	**0.0012**	0.0009	0.4169	0.0145	0.0266	0.0017	0.2452	76.49
Attention	553	**0.0129**	**0.0004**	**0.4691**	**0.0000**	0.0002	0.6248	0.0053	0.1200	0.0001	0.7150	0.0043	0.3090	0.0000	0.9083	67.33
Emotion	551	0.0040	0.0509	**0.2509**	**0.0000**	0.0026	0.1783	0.0098	0.0642	0.0071	0.0220	0.0097	0.1218	0.0004	0.5915	62.67
Episodic Memory	549	0.0000	0.9999	**0.1643**	**0.0000**	0.0015	0.3203	0.0095	0.1018	0.0012	0.3796	0.0058	0.4375	0.0010	0.4106	94.96
Working Memory	548	0.0033	0.0447	**0.4553**	**0.0000**	0.0016	0.2176	0.0104	0.0149	0.0001	0.7514	0.0063	0.1907	0.0015	0.2139	87.65
Balance	540	**0.0446**	**0.0000**	**0.1709**	**0.0000**	0.0031	0.1524	0.0069	0.1953	0.0032	0.1406	0.0102	0.1443	0.0092	0.0126	NA[b]
Total	519	**0.0166**	**0.0012**	**0.1577**	**0.0000**	0.0052	0.0765	0.0192	0.0080	0.0028	0.1881	0.0066	0.3986	0.0000	0.9415	68.32
Speed Composites																
General Ability	549	0.0089	0.0195	**0.3357**	**0.0000**	0.0066	0.0185	0.0029	0.4999	0.0017	0.2377	0.0121	0.0416	0.0002	0.7073	75.05
Abstraction	542	0.0000	0.9999	**0.2823**	**0.0000**	0.0000	0.9082	0.0069	0.1554	0.0011	0.3576	0.0082	0.1858	0.0023	0.1866	~100
Attention	550	0.0000	0.9999	**0.1432**	**0.0000**	0.0005	0.5519	0.0052	0.3330	0.0002	0.7443	0.0116	0.1081	0.0008	0.4570	NA[c]
Emotion	550	0.0053	0.0283	**0.2743**	**0.0000**	0.0002	0.7095	0.0119	0.0260	0.0031	0.1187	**0.0226**	**0.0014**	0.0007	0.4647	93.48
Episodic Memory	549	0.0076	0.0136	**0.2156**	**0.0000**	0.0001	0.7495	0.0093	0.0770	0.0000	0.9296	0.0130	0.0507	**0.0114**	**0.0038**	85.51
Working Memory	548	0.0000	0.9999	**0.2242**	**0.0000**	0.0000	0.9043	0.0039	0.3926	0.0017	0.2792	0.0138	0.0379	0.0003	0.6297	99.99
Motor	536	0.0000	0.9999	**0.1969**	**0.0000**	0.0017	0.2698	0.0040	0.4199	0.0052	0.0557	**0.0351**	**0.0001**	0.0001	0.7555	68.09
Total	519	0.0066	0.0261	**0.1987**	**0.0000**	0.0007	0.4909	0.0069	0.2054	0.0002	0.7424	**0.0338**	**0.0002**	0.0025	0.2016	69.58

Family-wise Bonferroni correction for 8 comparisons = 0.0063 ($\alpha = 0.05$, 2-tailed).
practice slope = observed slope – predicted group GAM fit baseline developmental slope from baseline age to 1-yr followup age.
site effects measured against UCSD.

[a] SES = socioeconomic status determined from the highest education of one parent.

[b] NA: The mean developmental slope was nonmonotonic, whereas the practice slope was monotonic, resulting in invalid ratios.

[c] NA: The mean developmental slope was nonmonotonic, whereas the average practice is negative and expected development is positive, resulting in invalid ratios.

rect responses for each of the three principles. If the participant does not achieve a principle within 48 trials, the test ends.

Matrix Analysis Test, a measure of abstraction and mental flexibility, is a multiple choice task in which the participant must conceptualize spatial, design, and numerical relations that range in difficulty from very easy to increasingly complex. The participant chooses a square that best fits in the missing space of a pattern. Patterns are made up of 2×2, 3×3, and 1×5 arrangements of squares. Each item has five response options.

Logical Reasoning, a measure of verbal intellectual ability, is a multiple-choice task in which the participant must complete verbal analogy problems.

2.3.2. Attention

The Continuous Performance task has two parts: one in which the participant must press the spacebar whenever lines form a complete number, and one whenever lines form a complete letter. Each part lasts 1.5 min. Each stimulus flashes for 300 ms followed by a blank page displayed for 700 ms, giving the participant 1 s to respond to each trial.

2.3.3. Emotion

For Emotion Recognition, participants view a series of 40 faces and indicate what emotion the face is showing: Happy, Sad, Angry, Scared, or No Feeling. There are 4 female faces for each emotion ($4 \times 5 = 20$) and 4 male faces for each emotion ($4 \times 5 = 20$).

Emotion Differentiation measures the ability to detect emotion intensity. The participant views pairs of faces and chooses the face showing greater intensity of emotion (anger, fear, happiness, sadness), or chooses a central button labeled Equal. The stimuli are created using software to morph faces into differing intensities of emotion. There are 36 trials, divided into happy, sad, angry, and fearful faces. Of the 36 trials, 4 show no emotional difference. The remaining 32 trials have emotion differentials in increments of 10% ranging from 10% to 60%, distributed more heavily toward 30% and 40% items. Trials are presented in random order, and the test is a forced-choice task with no time limit per trial.

2.3.4. Episodic memory

In the Face Memory test, participants are first shown 20 faces that they will be asked to identify later during immediate and delayed recognition trials. During immediate recall, participants view a series of 40 faces; 20 faces are targets for memory and 20 are distractors. Participants decide whether they had been previously shown the face by choosing one of four buttons, presented in a 4-point scale: "definitely yes," "probably yes," "probably no," and "definitely no" via the mouse. Delayed memory is tested approximately 25 min after immediate memory.

The Word Memory test is a verbal analogue to Face Memory and follows the same procedure for immediate and delayed recognition.

Visual Object Learning requires participants to view 10 three-dimensional Euclidean shapes that they will be asked to identify for immediate and delayed recognition in the same manner as Face Memory and Word Memory.

2.3.5. Working memory

Short Fractal N-back measures attention and working memory. Participants view fractal designs displayed on the computer screen and indicate the "target design.T̈here are three trial conditions. During the 0-back condition, the target design is designated before the trial and the participant responds each time they see it. For the 1-back and 2-back conditions, the target design is indicated by the repetition of a design, with the participants responding when they see a design that is the same as the design immediately preceding it (1-back condition) or is the same as the design immediately preceding the last design (2-back condition). In all trials, the participant has 2500 ms to respond.

2.3.6. Balance

Postural stability, measured with the modified Fregly-Graybiel Walk-a-Line ataxia test (Fregly et al., 1972; Sullivan et al., 2000), has 4 conditions and was conducted twice if the first trial was not completed perfectly. All trials were conducted with arms folded, eyes closed, and feet straight on a line on the floor: stand heel-to-toe for 60 s; stand on one and then the other foot for 30 s each; walk heel-to-toe for 10 steps.

2.3.7. General ability

Vocabulary from the WebCNP comprises five subtests, each containing 10 multiplechoice items with four response choices. The questions in each section are presented in order of increasing difficulty. A section is discontinued if the participant answers 5 questions incorrectly. Each subtest uses a different measure of verbal knowledge. In Part I, the participant chooses a word 'closest in meaning' to the target word. In Part II, the participant chooses the word that has a similar meaning to a bolded phrase within a sentence. In Part III, the participant selects the one word that is not a valid English word. In Part IV, the participant selects the word that is opposite in meaning to the target word. In Part V, the participant must choose the correct sentence based on contextual use of a target word.

The Wide Range Achievement Test-4 (WRAT4) assesses general ability in word reading and math calculation (blue form) (Wilkinson and Robertson, 2010); these scores were included in the General Ability composite.

2.3.8. Motor speed

This speed composite comprised response times from the Motor Praxis WebCNP test, which measures sensorimotor ability by having the participant use the mouse to click on a shrinking box when it moves to a new position on the screen. This test screens a participant's dexterity, an essential ability to perform the WebCNP tests.

2.4. Statistical analysis

The analyses focused on neuropsychological data acquired from the 568 adolescents in the no/low-drinking group that continued to meet criteria across time. This set of analyses assessed the contributions of practice, age, sex, ethnicity, SES, examination site, and family history on change scores.

Performance measures of each participant at both test times were expressed as Z-scores calculated from all available baseline scores of the no/low group. Slopes of each measure for each participant were the Time 2 score minus the Time 1 score divided by the test interval, which was approximately 1 year. Prior testing experience, i.e., practice, was operationalized as the observed change over 1 year minus the predicted change over 1 year, determined from the cross-sectional baseline age regression function (Fig. 1). For example, a 16.4 year old boy at baseline achieving a score of −0.4 Z and a follow-up score of + 0.3 Z would have an observed slope = 0.7 Z/year. If the average cross-sectional difference between 16.4 year-old and 17.4 year-old boys is +0.1 Z, the practice slope of the participant at the second test would be 0.6 Z/year. Conversely, the developmental effect was the observed change minus the practice score (0.7 Z − 0.6 Z = 0.1). Estimation of the contribution of practice relative to expected development on the observed slopes indicated that the majority of the change over the year was attributable to practice rather than development for almost all measures (Table 2, last column). Therefore, practice effect was the primary unit of analysis for all composites and variables in the following analyses.

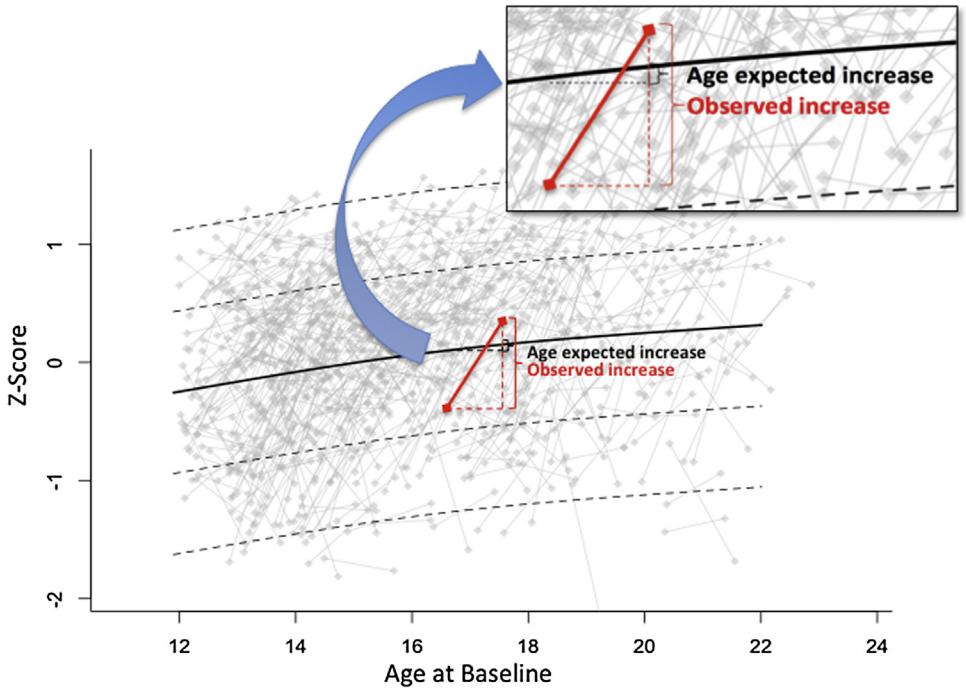

Fig. 1. Individual baseline and follow-up Z-scores for Abstraction Accuracy composite (gray). Predicted age-dependent improvement determined from the cross-sectional baseline age regression function (solid black line) with ± 1 and 2 SD (dashed black lines). Cutout: example of a single subject age 16.4 years at baseline and 17.4 years at followup (red). At baseline his Z-score was −0.4 SD and at follow-up it was +0.3 SD = 0.7 SD "Observed increase" (red). The expected age-dependent improvement determined from the baseline age regression function was 0.1 SD "Age expected increase" (black). Thus, his practice effect was 0.7 −0.1 = 0.6 SD. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The primary analysis tools were the General Additive Model (GAM) (Hastie and Tibshirani, 1986, 1990; Wood, 2006, 2011) and analysis of variance (ANOVA) from the "mgcv" package in R (Wood, 2006) Version 3.1.0 [http://www.r-project.org/], testing for the predictive value of the main effect of age with selected covariates. The GAM tested the predictive value of age and 6 covariates (with respect to each subject, *i*)—baseline performance (baseline $_i$), SES (SES_i), ethnicity (ethnicity $_i$), sex (sex $_i$), examination site (site $_i$), and family history (FH_i)—on each performance score [i.e., practice effects = (observed slope − predicted slope)/interval].

$$GAM: \textit{composite practice slope}_i \sim \beta_0 + S_1(\text{baseline age}_i) + \beta_2 \text{baseline score}_i + \beta_3 SES_i + \beta_4 \text{ethnicity}_i + \beta_5 \text{sex}_i + \beta_6 \text{site}_i + \beta_7 FH_i$$

Age was allowed to be a nonlinear smooth effect, implemented via thin plate splines $S_1(age)$ with 3 knots (Wood, 2003). Roughness penalties for the smooth effects were estimated using generalized cross validation (Wood, 2004).

Many scores were modulated by several or all covariates. Contributions of covariates were considered significant with $p \leq 0.0063$, reflecting family-wise Bonferroni correction for 8 comparisons with $\alpha = 0.05$, 2-tailed. The sample sizes varied slightly across models tested (noted in the results tables) because not all participants had data for all test measures.

3. Results

3.1. Performance change over 1 year

With a few exceptions, the average extent of improvement in performance over 1 year was primarily attributable to prior testing experience with little contribution from predicted developmental effects (Table 2, last column; Figs. 2 and 3). Note that the cross-sectional age regressions produced from the GAM at follow-up (orange lines) are similar to those at baseline (black lines), only shifted in the positive direction reflecting the practice effect and to the right by one year. The group mean practice slope, expressed as the percent of the group mean observed slope, estimated that for 13 composites, approximately 63% to nearly 100% was attributable to previous experience with the tests. In 2 other instances, these ratios were invalid because either the developmental or practice slope was nonmonotonic while the complementary slope was monotonic. The measure showing the least practice effect (29.2%) was the General Ability Accuracy composite, suggesting a greater contribution from development to performance change on this composite score than on any other test composite score.

3.2. Baseline test performance

In the full general additive model the amount of variance that was attributable to baseline test performance was highly significant for all 16 composite scores and ranged from a high of 47% for the Attention Accuracy composite to a low of 5% for the General Ability Accuracy composite (Table 2 and Fig. 4). For all 16 composite scores, participants who *scored lowest* at baseline showed the *greatest gain* over the 1-year interval.

3.3. Age

Age made small (1.5%–4.6%) but significant contributions to the variance of 4 accuracy composite scores: Abstraction, Attention, Balance, and Total (Table 2, Fig. 5). In general, the younger participants showed greater gain than their older counterparts. This age effect is consistent with the low baseline score effect because lower scores were typically achieved by younger participants.

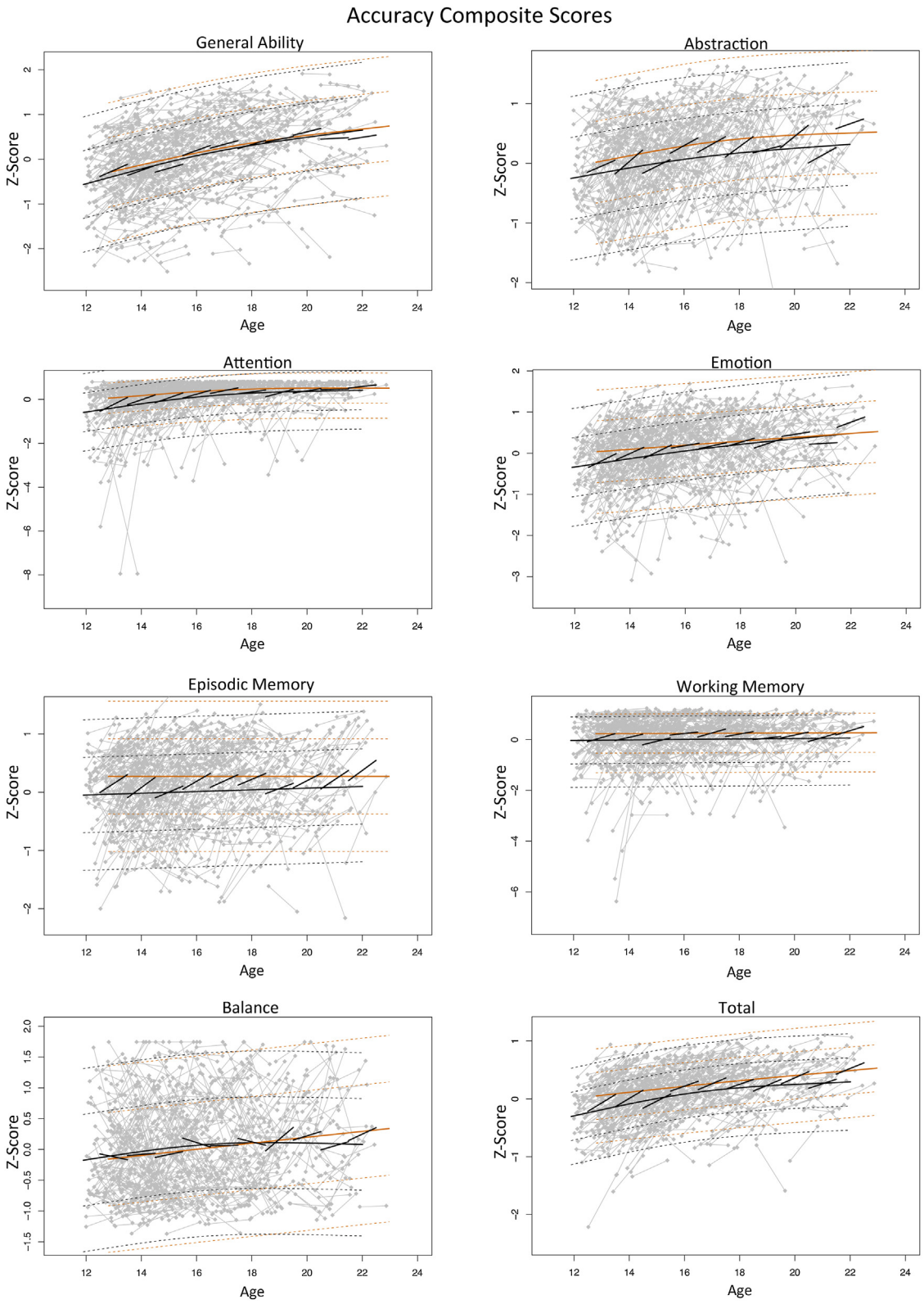

Fig. 2. Individual baseline and follow-up Z-scores for the Accuracy composite scores (gray). Predicted age-dependent improvement determined from the cross-sectional baseline age regression function is plotted in solid black line with ± 1 and 2 SD as dashed black lines. The predicted age-dependent improvement determined from the cross-sectional follow-up age regression function is plotted in orange. Note the baseline and follow-up cross-sectional functions are similar with the latter being shifted positively and beginning and ending a year later than the baseline function. The individual black lines at yearly intervals are the average baseline and follow-up Z-scores for each year (slope of improvement) for all 12 year olds, 13 year olds, etc. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

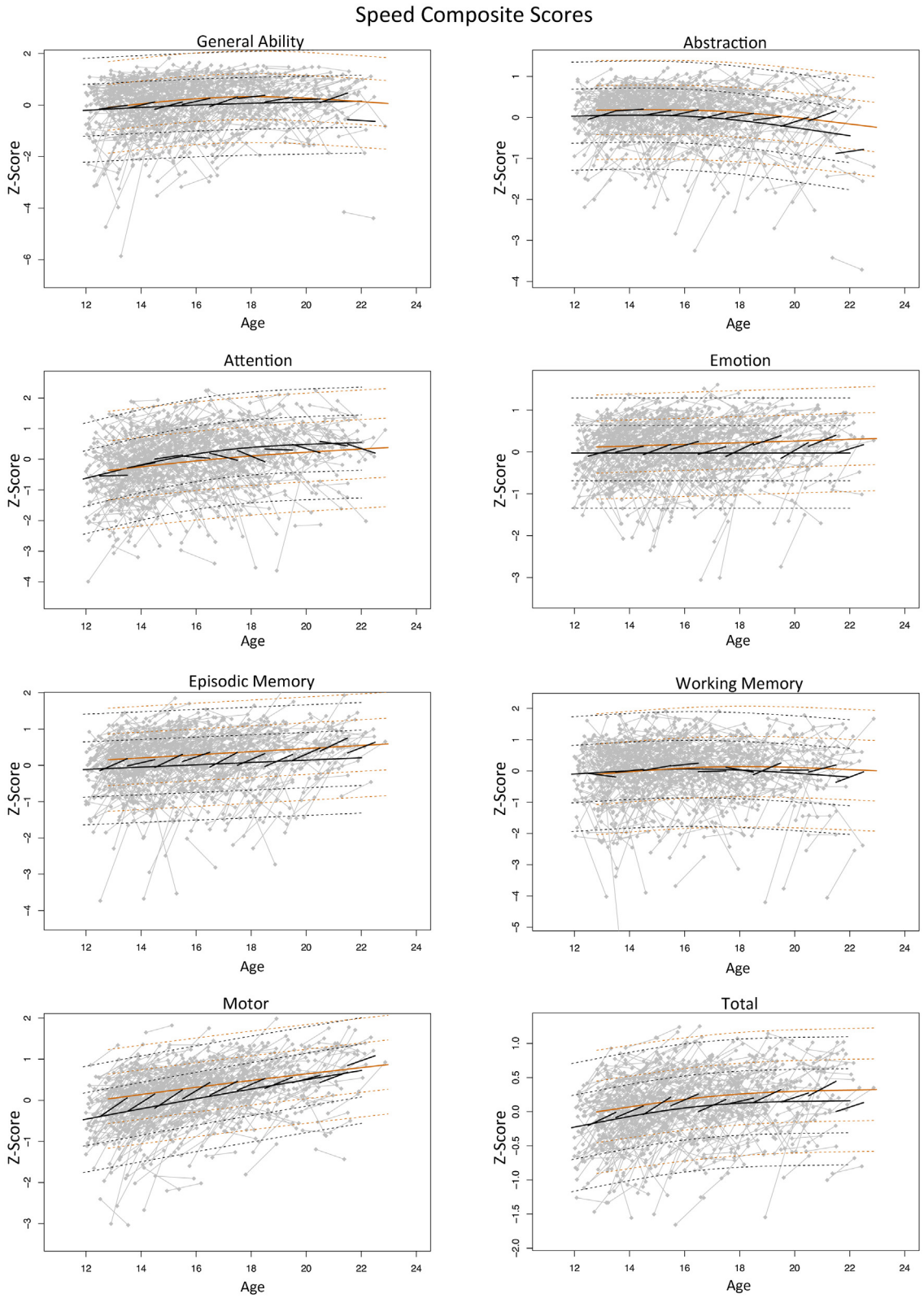

Fig. 3. Individual baseline and follow-up Z-scores for Speed Composite Scores (gray). Predicted age-dependent improvement determined from the cross-sectional baseline age regression function is plotted in solid black line with ± 1 and 2 SD as dashed black lines. The predicted age-dependent improvement determined from the cross-sectional follow-up age regression function is plotted in orange. Note the baseline and follow-up cross-sectional functions are similar with the latter being shifted positively and beginning and ending a year later than the baseline function. The individual black lines at yearly intervals are the average baseline and follow-up values for each year (slope of improvement) for all 12 year olds, 13 year olds, etc. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

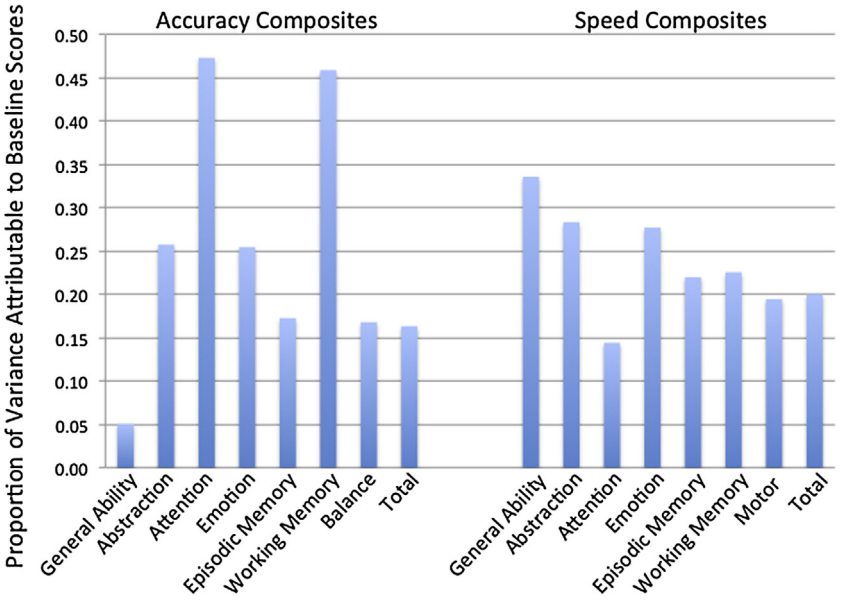

Fig. 4. Bar graph presentation of the proportion of variance accounted for by the baseline performance in the GAM predicting practice effect for the Accuracy and Speed Composites.
GAM: composite practice slope$_i \sim \beta_0 + S_1$(baseline age $_i$) + $\beta_2$baseline score $_i$ + β_3SES$_i$ + $\beta_4$ethnicity $_i$ + $\beta_5$sex $_i$ + $\beta_6$site $_i$ + β_7FH$_i$.

3.4. Socioeconomic status (SES)

SES, in terms of the highest education level achieved by a parent, was significant and accounted for ~1.7% of the variance for General Ability and Abstraction Accuracy composites, where the higher the education, the larger the practice effect (Table 2).

3.5. Ethnicity

Ethnicity accounted for small (1.3% to 3.1%) but statistically significant proportions of variance for 4 accuracy scores. Specifically, African Americans and Asians showed less of a practice effect than Caucasians on the Abstraction Accuracy composite (Table 2). Further, the practice effect exhibited by African Americans was smaller than that observed in Caucasians for 3 additional accuracy composites (Episodic Memory, Working Memory, and Total), suggestive of similar performance over the test interval.

3.6. Sex

Regarding sex, no differences were detected between male and female participants with respect to gains from prior test experience (Table 2).

3.7. Examination site

The UCSD site participants showed a greater advantage from prior test experience than the remaining 4 consortium sites on speeded performance for the Emotion, Motor, and Total Speed composite scores accounting for 2–3% of the variance (Table 2).

3.8. Family history of alcohol or substance abuse

Family history exerted a statistically significant influence on only one measure; here, family history positive participants exhibited greater gain from prior experience with testing material in the Episodic Memory Speed composite than did family history negative participants, although accounting for only 1.1% of the variance.

3.9. Proportion of participants with performance gains or losses

Chi-square tests of the proportion of participants who improved relative to those whose scores remained the same or declined over the 1-year retest interval failed to identify differences related to sex or ethnicity for any test composite. By contrast, the proportion of participants whose scores improved over the interval was significantly higher than those whose scores were the same or lower on 6 accuracy and 6 speed measures, again indicative of benefit from prior experience with the tests (Table 3).

4. Discussion

This longitudinal analysis addressed two study aims. The first aim was to distinguish and quantify developmental effects from prior testing experience effects (commonly called practice effects) on cognitive and motor measures over a 1-year interval. Here, we found that previous experience with the test materials and procedures accounted for a substantial amount of variance of change scores in accuracy and speed performance on most aggregated functions examined. The second aim estimated the variance accounted for by age, baseline performance, sex, SES, ethnicity, family history of alcohol or other substance use, and examination site to performance change and found that baseline score accounted for a greater amount of variance in change scores (ranging from ~5% to 47% for accuracy and ~14% to 34% for speed change scores) than any other factor examined. These results are expanded next.

The most salient finding of this study was that prior experience with the test material and procedures, also considered practice effects, was the overwhelmingly strongest factor affecting both accuracy and speed change scores, despite a 1-year test interval. To estimate practice effects, we used a version of the twice-minus-once-tested method (Anderson et al., 2001; Salthouse, 2015) by comparing longitudinally-determined change over the 1-year test interval and change estimated from cross-sectionally-determined developmental trajectories. Some studies have shown that the greatest practice effect occurs between the first and second testing (Ibrahim et al., 2015; Waber et al., 2012) but diminishes at later

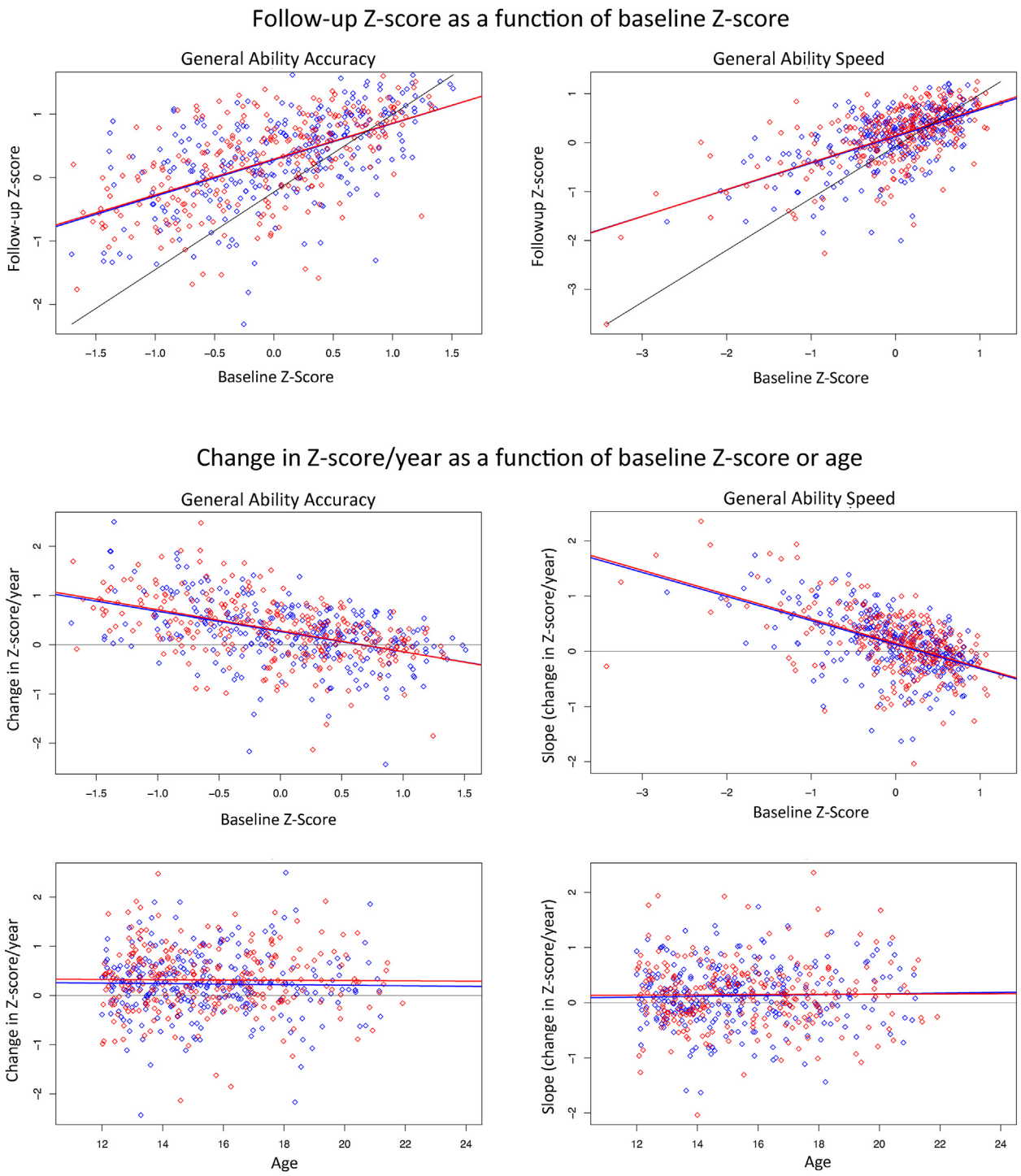

Fig. 5. Examples of the relation of baseline and follow-up data for the General Ability Accuracy (left) and Speed (right) Composites (boys in blue; girls in red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Top: Follow-up Z-score (y) as a function of Baseline Z-score (x); isoline, i.e., no difference (gray), boy (blue) and girl (red) linear regression lines.
Middle: Slope (change in 1 year) in Z-score/year as a function of Baseline Z-Score; isoline, i.e., no difference (gray), boy (blue) and girl (red) linear regression lines.
Bottom: Slope (change in 1 year) in Z-score/year as a function of Age; isoline, i.e., no difference (gray), boy (blue) and girl (red) linear regression lines.

repeated testing, observed in studies of pre-adolescence and adolescence development (Falleti et al., 2006; Thomas et al., 2016) and of healthy adults (e.g., Anderson et al., 2001; Ivnik et al., 2000; Nettelbeck and Wilson, 1985; Salthouse and Tucker-Drob, 2008). Whether developmental change ultimately overrides practice effects with later follow-up testing remains to be examined.

Not all participants improved over the 1-year interval, but χ^2 tests indicated that more participants gained than lost points over the 1-year. The general trend in the present study was that the younger the participant at baseline, the greater the gain at the 1-year follow-up, especially on Abstraction, Attention, and Balance accuracy. A factor potentially contributing to this age effect

Table 3
Proportion of participants whose scores improved over 1 year vs. declined or remained at baseline level.

Composite	Sex		Ethnicity		Family History		Positive	Negative			
	$\chi 2$	p	$\chi 2$	p	$\chi 2$	p	N	N	$\chi 2$	p	Direction
Accuracy Composites											
General Ability	0.3020	0.5827	1.8207	0.4024	0.0000	1.0000	359	277	5.0524	0.0246	p > n
Abstraction	0.2806	0.5963	1.6759	0.4326	0.0000	1.0000	416	216	31.8123	**0.0000**	**p > n**
Attention	1.6138	0.2040	1.0978	0.5776	0.8096	0.3682	390	256	13.6327	**0.0002**	**p > n**
Emotion	1.9709	0.1604	2.8611	0.2392	1.4513	0.2283	372	271	7.6686	**0.0056**	**p > n**
Episodic Memory	0.1082	0.7422	1.3535	0.5083	0.0000	1.0000	464	177	66.7009	**0.0000**	**p > n**
Working Memory	0.9413	0.3320	3.0795	0.2144	0.1218	0.7271	387	253	13.7633	**0.0002**	**p > n**
Balance	0.0155	0.9008	0.4385	0.8031	0.0417	0.8381	302	329	0.4955	0.4815	p < n
Total	0.0432	0.8353	0.8946	0.6394	0.0000	1.0000	442	163	66.9731	**0.0000**	**p > n**
Speed Composites											
General Ability	0.0022	0.9624	1.3484	0.5096	0.0005	0.9829	377	264	9.6860	**0.0019**	**p > n**
Abstraction	3.8785	0.0489	1.9744	0.3726	0.0219	0.8823	382	250	13.5177	**0.0002**	**p > n**
Attention	0.0011	0.9730	0.7426	0.6898	0.1259	0.7227	273	370	7.0580	0.0079	p < n
Emotion	1.1767	0.2780	0.9811	0.6123	0.0220	0.8821	422	219	32.3245	**0.0000**	**p > n**
Episodic Memory	0.5667	0.4516	6.7139	0.0348	0.0000	1.0000	451	190	54.5878	**0.0000**	**p > n**
Working Memory	0.0565	0.8121	1.7275	0.4216	0.9688	0.3250	342	298	1.3798	0.2401	p > n
Motor	0.7884	0.3746	0.2895	0.8653	0.5489	0.4588	441	184	54.3160	**0.0000**	**p > n**
Total	0.6824	0.4088	1.0231	0.5996	1.4056	0.2358	394	209	28.4378	**0.0000**	**p > n**

Family-wise Bonferroni correction for 8 comparisons = 0.0063 ($\alpha = 0.05$, 2-tailed).

was ceiling performance, which was more likely to occur in older than younger adolescents, thereby leaving less range for change in the older than younger participants. A similar age-related gain in response speed and accuracy was reported in youth age 9–12 years that stabilized by age 15 years (Williams et al., 2016). Also consistent with the current pattern are other studies that sought and reported practice effects in some but not all functions examined. For example, a study aimed as distinguishing between test-retest reliability and developmental changes over 5 years in youth age 8–12 years at baseline found that of the 19 neuropsychological tests given, 1) the greatest gain was typically observed in the younger participants, 2) performance variance stabilized after 2 or 3 test years, and 3) practice effects differed by test: the greatest practice effects occurred on the Grooved Pegboard test, moderate practice effects occurred on figure matching and symbol search, and no measurable practice effects occurred on Digit Span (Slade et al., 2008).

Although the overall trend was toward improvement in scores, there was substantial variability in the degree of the influence of testing experience. In particular, on 13 of the 14 different functional domains examined, 63% to nearly 100% of the change was attributable to practice rather than development. The exception was for General Ability accuracy, for which only 29% of the improvement was due to prior experience. Further, although baseline performance was a strong predictor of gain in all 16 composite scores, when entered as a covariate in the general additive model the least affected was General Ability accuracy, accounting for only 5% of the variance compared with 14%–47% for the remaining scores.

Sex, site, and ethnicity contributed little to nothing to the change variance in the present study. Girls and boys did not differ at any age in gain over a year. Of the 5 sites, UCSD participants showed greater gain than the other 4 sites on 3 speed measures. The small but statistically significant ethnicity effect indicated greater performance consistency, that is, less of a prior experience effect, on a few measures by African Americans and Asians.

Given the vast number of tests, measures, and stimuli in the test battery, it is unlikely that specific stimuli or responses were recalled, although items could be more readily recognized even after 1 year. Thus, episodic recognition or recall of specific test items could only partially account for the large practice effects. Another factor possibly contributing to practice is procedural learning. This possibility is supported anecdotally by a lead test administrator [D.P.], who noted that, unlike at baseline, at follow-up examinations participants typically needed test instructions to be read only once before starting a task and required only one practice trial to be introduced to a task.

5. Conclusion

Factors that made substantial contributions to change in performance over 1 year were prior experience with the testing materials and procedures, baseline performance level, and age, where younger boys and girls who achieved lower scores at baseline showed the greatest gain at retest. SES, ethnicity, and testing site made significant but trivial contributions to performance change, whereas both sexes showed similar levels of improvement. Estimates of practice, also considered a form of learning, may be a relevant metric for detecting interim influences on behavior.

Studies of adolescent populations that are likely to have significant developmental gains in performance over a few years face a trade off. On the one hand, the more subjects there are at a common age, the more power is available to detect preexisting or external factors that interact with development. On the other hand, if all participants in longitudinal study are essentially the same age, there is little opportunity to disentangle practice from development effects. In the current study, the three-age cohort, accelerated longitudinal design allowed for construction of cross-sectional developmental predictions across ten years that were consistent at baseline and one-year follow-up, with the exception of the quantum increase in performance attributable to prior experience, i.e., practice effects. These considerations and trade-offs should inform other large-scale studies with respect to including a broad enough age range at study initiation to enable estimation of cross-sectional age-performance trajectories independent of practice effects. Indeed, Salthouse (Salthouse, 2015) warned that "prior experience with the tests may not only lead to underestimates of cognitive declines in adulthood, but also to overestimates of cognitive gains in childhood" (page 1269), indicating the usefulness of the twice-minus-once-tested method to distinguish practice effects from ontological changes associated with development or senescence.

Conflict of interest

None.

Submission declaration and verification

We verify that the work described in the submitted manuscript has not been published previously.

References

Akshoomoff, N., Newman, E., Thompson, W.K., McCabe, C., Bloss, C.S., Chang, L., Amaral, D.G., Casey, B.J., Ernst, T.M., Frazier, J.A., Gruen, J.R., Kaufmann, W.E., Kenet, T., Kennedy, D.N., Libiger, O., Mostofsky, S., Murray, S.S., Sowell, E.R., Schork, N., Dale, A.M., Jernigan, T.L., 2014. The NIH Toolbox Cognition Battery: results from a large normative developmental sample (PING). Neuropsychology 28, 1–10.

Anderson, M., Reid, C., Nelson, J., 2001. Developmental changes in inspection time: what a difference a year makes. Intelligence 29, 475–486.

Brown, S.A., Myers, M.G., Lippke, L., Tapert, S.F., Stewart, D.G., Vik, P.W., 1998. Psychometric evaluation of the Customary Drinking and Drug Use Record (CDDR): A measure of adolescent alcohol and drug involvement. J. Stud. Alcohol 59 (4), 427–438.

Brown, S.A., Brumback, T., Tomlinson, K., Cummins, K., Thompson, W.K., Nagel, B.J., De Bellis, M.D., Hooper, S.R., Clark, D.B., Chung, T., Hasler, B.P., Colrain, I.M., Baker, F.C., Prouty, D., Pfefferbaum, A., Sullivan, E.V., Pohl, K.M., Rohlfing, T., Nichols, B.N., Chu, W., Tapert, S.F., 2015. The National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA): A multi-site study of adolescent development and substance use. J. Stud. Alcohol Drugs 76, 895–908.

Calamia, M., Markon, K., Tranel, D., 2012. Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. Clin. Neuropsychol. 26 (4), 543–570, http://dx.doi.org/10.1080/13854046.2012.680913.

Cromer, J.A., Schembri, A.J., Harel, B.T., Maruff, P., 2015. The nature and rate of cognitive maturation from late childhood to adulthood. Front. Psychol. 6, 704, http://dx.doi.org/10.3389/fpsyg.2015.00704.

Falleti, M.G., Maruff, P., Collie, A., Darby, D.G., 2006. Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals. J. Clin. Exp. Neuropsychol. 28 (7), 1095–1112, http://dx.doi.org/10.1080/13803390500205718.

Fregly, A.R., Graybiel, A., Smith, M.S., 1972. Walk on floor eyes closed (WOFEC): a new addition to an ataxia test battery. Aerosp. Med. 43 (4), 395–399.

Giedd, J.N., Raznahan, A., Alexander-Bloch, A., Schmitt, E., Gogtay, N., Rapoport, J.L., 2014. Child psychiatry branch of the National Institute of Mental Health longitudinal structural magnetic resonance imaging study of human brain development. Neuropsychopharmacology 40 (1), 43–49, http://dx.doi.org/10.1038/npp.2014.236, npp2014236 [pii].

Gur, R.C., Richard, J., Hughett, P., Calkins, M.E., Macy, L., Bilker, W.B., Brensinger, C., Gur, R.E., 2010. A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: standardization and initial construct validation. J. Neurosci. Methods 187, 254–262.

Gur, R.C., Richard, J., Calkins, M.E., Chiavacci, R., Hansen, J.A., Bilker, W.B., Loughead, J., Connolly, J.J., Qiu, H., Mentch, F.D., Abou-Sleiman, P.M., Hakonarson, H., Gur, R.E., 2012. Age group and sex differences in performance on a computerized neurocognitive battery in children age 8–21. Neuropsychology 26, 251–265.

Hastie, T., Tibshirani, R., 1986. Generalized additive models (with DIscussion). Statistical Science 1, 297–318.

Hastie, T., Tibshirani, R., 1990. Exploring the nature of covariate effects in the proportional hazards model. Biometrics 46 (4), 1005–1016.

Ibrahim, I., Tobar, S., Elassy, M., Mansour, H., Chen, K., Wood, J., Gur, R.C., Gur, R.E., El Bahaei, W., Nimgaonkar, V., 2015. Practice effects distort translational validity estimates for a Neurocognitive Battery. J. Clin. Exp. Neuropsychol. 37, 530–537.

Ivnik, R.J., Smith, G.E., Petersen, R.C., Boeve, B.F., Kokmen, E., Tangalos, E.G., 2000. Diagnostic accuracy of four approaches to interpreting neuropsychological test data. Neuropsychology 14 (2), 163–177.

Jernigan, T.L., Brown, T.T., Hagler Jr., D.J., Akshoomoff, N., Bartsch, H., Newman, E., Thompson, W.K., Bloss, C.S., Murray, S.S., Schork, N., Kennedy, D.N., Kuperman, J.M., McCabe, C., Chung, Y., Libiger, O., Maddox, M., Casey, B.J., Chang, L., Ernst, T.M., Frazier, J.A., Gruen, J.R., Sowell, E.R., Kenet, T., Kaufmann, W.E., Mostofsky, S., Amaral, D.G., Dale, A.M., 2016. Pediatric Imaging N, Genetics S. The Pediatric Imaging, Neurocognition, and Genetics (PING) Data Repository. Neuroimage 124, 1149–1154.

Nettelbeck, T., Wilson, C., 1985. A cross-sequential analysis of developmental differences in speed of visual information processing. J. Exp. Child Psychol. 40 (1), 1–22.

Nichols, B.N., Pohl, K.M., 2015. Neuroinformatics software applications supporting electronic data capture, management, and sharing for the neuroimaging community. Neuropsychol. Rev. 25 (3), 356–368, http://dx.doi.org/10.1007/s11065-015-9293-x.

Roalf, D.R., Gur, R.C., Almasy, L., Richard, J., Gallagher, R.S., Prasad, K., Wood, J., Pogue-Geile, M.F., Nimgaonkar, V.L., Gur, R.E., 2013. Neurocognitive performance stability in a multiplex multigenerational study of schizophrenia. Schizophr. Bull. 39, 1008–1017.

Rohlfing, T., Cummins, K., Henthorn, T., Chu, W., Nichols, B.N., 2014. N-CANDA data integration: anatomy of an asynchronous infrastructure for multi-site, multi-instrument longitudinal data capture. J. Am. Med. Inform. Assoc. 21 (4), 758–762, http://dx.doi.org/10.1136/amiajnl-2013-002367, amiajnl-2013-002367 [pii].

Salthouse, T.A., Tucker-Drob, E.M., 2008. Implications of short-term retest effects for the interpretation of longitudinal change. Neuropsychology 22 (6), 800–811, http://dx.doi.org/10.1037/a0013091.

Salthouse, T.A., 2015. Test experience effects in longitudinal comparisons of adult cognitive functioning. Dev. Psychol. 51 (9), 1262–1270, http://dx.doi.org/10.1037/dev0000030.

Satterthwaite, T.D., Connolly, J.J., Ruparel, K., Calkins, M.E., Jackson, C., Elliott, M.A., Roalf, D.R., Ryan Hopsona, K.P., Behr, M., Qiu, H., Mentch, F.D., Chiavacci, R., Sleiman, P.M., Gur, R.C., Hakonarson, H., Gur, R.E., 2016. The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth. Neuroimage 124, 1115–1119.

Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Buchel, C., Conrod, P.J., Dalley, J.W., Flor, H., Gallinat, J., Garavan, H., Heinz, A., Itterman, B., Lathrop, M., Mallik, C., Mann, K., Martinot, J.L., Paus, T., Poline, J.B., Robbins, T.W., Rietschel, M., Reed, L., Smolka, M., Spanagel, R., Speiser, C., Stephens, D.N., Strohle, A., Struve, M., consortium, I., 2010. The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. Mol. Psychiatry 15, 1128–1139.

Sirois, P.A., Posner, M., Stehbens, J.A., Loveland, K.A., Nichols, S., Donfield, S.M., Bell, T.S., Hill, S.D., Amodei, N., Hemophilia, G., Development, S., 2002. Quantifying practice effects in longitudinal research with the WISC-R and WAIS-R: a study of children and adolescents with hemophilia and male siblings without hemophilia. J. Pediatr. Psychol. 27, 121–131.

Slade, P.D., Townes, B.D., Rosenbaum, G., Martins, I.P., Luis, H., Bernardo, M., Martin, M.D., Derouen, T.A., 2008. The serial use of child neurocognitive tests: development versus practice effects. Psychol. Assess. 20, 361–369.

Sullivan, E.V., Shear, P.K., Zipursky, R.B., Sagar, H.J., Pfefferbaum, A., 1994. A deficit profile of executive, memory, and motor functions in schizophrenia. Biol. Psychiatry 36 (10), 641–653.

Sullivan, E.V., Rosenbloom, M.J., Pfefferbaum, A., 2000. Pattern of motor and cognitive deficits in detoxified alcoholic men. Alcohol.: Clin. Exp. Res. 24 (5), 611–621.

Sullivan, E.V., Brumback, T., Tapert, S.F., Fama, R., Prouty, D., Brown, S.A., Cummins, K., Thompson, W.K., Colrain, I.M., Baker, F.C., De Bellis, M.D., Hooper, S.R., Clark, D.B., Chung, T., Nagel, B.J., Nichols, B.N., Rohlfing, T., Chu, W., Pohl, K.M., Pfefferbaum, A., 2016. Cognitive, emotion control, and motor performance of adolescents in the NCANDA study: Contributions from alcohol consumption, age, sex, ethnicity, and family history of addiction. Neuropsychology 30, 449–473.

Thomas, J.M., Higgs, S., Dourish, C.T., 2016. Test-retest reliability and effects of repeated testing and satiety on performance of an Emotional Test Battery. J. Clin. Exp. Neuropsychol. 38 (4), 416–433, http://dx.doi.org/10.1080/13803395.2015.1121969.

Waber, D.P., Forbes, P.W., Almli, C.R., Blood, E.A., Brain Development Cooperative G, 2012. Four-year longitudinal performance of a population-based sample of healthy children on a neuropsychological battery: the NIH MRI study of normal brain development. J. Int. Neuropsychol. Soc. 18 (2), 179–190, http://dx.doi.org/10.1017/S1355617711001536.

Wilkinson, G.S., Robertson, G.J., 2010. Wide Range Achievement Test 4 (WRAT4).

Williams, J., Crowe, L.M., Dooley, J., Collie, A., Davis, G., McCrory, P., Clausen, H., Maddocks, D., Anderson, V., 2016. Developmental trajectory of information-processing skills in children: computer-based assessment. Appl. Neuropsychol. Child. 5, 35–43.

Wood, S.N., 2003. Thin-plate regression splines. J. R. Stat. Soc. B 65, 95–114.

Wood, S.N., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. J. Am. Stat. Assoc. 99, 673–686.

Wood, S.N., 2006. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. Biometrics 62 (4), 1025–1036, http://dx.doi.org/10.1111/j.1541-0420.2006.00574.x, BIOM574 [pii].

Wood, S.N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear modelsFast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. B 73 (1), 3–36.