Sullivan *et al.*
BMC Medical Research Methodology (2022) 22:177
https://doi.org/10.1186/s12874-022-01606-9

BMC Medical Research Methodology

RESEARCH

Open Access

Prior test experience confounds longitudinal tracking of adolescent cognitive and motor development

Edith V. Sullivan[1*], Wesley K. Thompson[2], Ty Brumback[3], Devin Prouty[4], Susan F. Tapert[5], Sandra A. Brown[5], Michael D. De Bellis[6], Kate B. Nooner[7], Fiona C. Baker[4], Ian M. Colrain[4], Duncan B. Clark[8], Bonnie J. Nagel[9], Kilian M. Pohl[1,4] and Adolf Pfefferbaum[1,4]

Abstract

Background: Accurate measurement of trajectories in longitudinal studies, considered the gold standard method for tracking functional growth during adolescence, decline in aging, and change after head injury, is subject to confounding by testing experience.

Methods: We measured change in cognitive and motor abilities over four test sessions (baseline and three annual assessments) in 154 male and 165 female participants (baseline age 12–21 years) from the National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA) study. At each of the four test sessions, these participants were given a test battery using computerized administration and traditional pencil and paper tests that yielded accuracy and speed measures for multiple component cognitive (Abstraction, Attention, Emotion, Episodic memory, Working memory, and General Ability) and motor (Ataxia and Speed) functions. The analysis aim was to dissociate neurodevelopment from testing experience by using an adaptation of the twice-minus-once tested method, which calculated the difference between longitudinal change (comprising developmental plus practice effects) and practice-free initial cross-sectional performance for each consecutive pairs of test sessions. Accordingly, the first set of analyses quantified the effects of *learning* (i.e., prior test experience) on accuracy and after speed domain scores. Then *developmental* effects were determined for each domain for accuracy and speed having removed the measured learning effects.

Results: The greatest gains in performance occurred between the first and second sessions, especially in younger participants, regardless of sex, but practice gains continued to accrue thereafter for several functions. For all 8 accuracy composite scores, the developmental effect after accounting for learning was significant across age and was adequately described by linear fits. The learning-adjusted developmental effects for speed were adequately described by linear fits for Abstraction, Emotion, Episodic Memory, General Ability, and Motor scores, although a nonlinear fit was better for Attention, Working Memory, and Average Speed scores.

Conclusion: Thus, what appeared as accelerated cognitive and motor development was, in most cases, attributable to learning. Recognition of the substantial influence of prior testing experience is critical for accurate characterization of normal development and for developing norms for clinical neuropsychological investigations of conditions affecting the brain.

Keywords: Longitudinal, Practice effects, Development, Cognition, Motor

*Correspondence: edie@stanford.edu

[1] Department of Psychiatry & Behavioral Sciences, Stanford University School of Medicine (MC5723), 401 Quarry Road, Stanford, CA 94305-5723, USA
Full list of author information is available at the end of the article

Background

Longitudinal studies are considered the gold standard protocol for tracking developmental and involutional changes. By nature, longitudinal assessment requires repeated examination, ideally employing the same procedures and test materials throughout the study [1, 2]. Some assessment classes are relatively robust to repeated testing, such as measurement using structural brain imaging with MRI, somatic size, or blood chemistry panels. Even such practice-free retesting is subject to measurement drift, which can be estimated with longitudinally acquired, control data to be used as correction factors (e.g., [3, 4]). By contrast, longitudinal cognitive assessment has the intrinsic problem of *prior test experience* [5–7], also considered "practice" or "learning" [8], even when the retest interval spans one [9] to two [10–13] years. Thus, any longitudinal study purporting to track, quantify, and infer cognitive change as development, maturation, or decline is confounded by prior testing experience that requires quantification (review, [14]).

Many studies that have considered practice effects have focused on adult aging to senescence [1, 8, 15] or on repeated testing necessary in clinical settings to track the progression of CNS injury due to accident, stroke, or dementia [16–19] or recovery with treatment [20] or time [21]. Indeed, practice effects have been speculated to minimize age-related declines in older people [22, 23] and have proved useful in predicting cognitive decline or stability in patients with amnestic Mild Cognitive Impairment (aMCI). Specifically, patients with aMCI, whose cognitive scores improved with repeated testing separated by one week, showed a relatively stable disease course over one year, whereas those who showed minimal improvement between the weekly testing evidenced substantial decline over one year [24]. Thus as further emphasized by Duff and colleagues in the title of their paper [8], practice effects can be considered a "unique cognitive variable".

A growing number of large-scale, longitudinal studies have been initiated to measure cognitive, motor, and emotional development from later childhood through young adulthood (reviewed in [25]). Among them, National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA) [26] with its cohort sequential design (described below) is uniquely positioned to measure practice effects and to dissociate them from developmental trajectories, modeling of which is the intent of all of these projects.

For decades, sex has been recognized as a significant moderator variable in studies of development of cognitive and motor processes (reviewed in [27, 28]). As we noted in our earlier paper [29], sex differences are associated with neuropsychological test performance during normal development and require consideration when assessing developmental trajectories of cognitive and motor functions (e.g., [30–32]). Typically, girls undergo sexual maturity earlier than boys (e.g., [33, 34]) and advance earlier than boys in language skills [35], use of semantic knowledge [36], facial emotion recognition and discrimination [37, 38], and components of episodic memory [37, 39]. By contrast, boys develop earlier than girls in mental rotation appreciation [40, 41], fine motor control (but see[42]) [43, 44], and physical strength (e.g., [45, 46]). Many sex-related differences identified are relevant to the tests used in the current study, girls tend to develop language skills earlier than boys, whereas boys develop spatial skills earlier than girls [37]. Into adolescence, as groups, girls excel on tests of memory and social cognition, whereas boys excel on tests of spatial processing and motor speed (reviewed in [27]).

Several methods have been proposed to preclude or minimize learning effects in longitudinal studies of cognitive and motor performance or, alternatively, to dissociate testing-experience learning from development. As summarized by Salthouse [5, 6] and McArdle [2], some methods, other than simply using cross-sectional protocols, include use of different test forms, staggering baseline testing, and the "twice-minus-once-tested" method [1, 5, 47]. The last approach has proved useful in studies employing an *accelerated longitudinal* design, also known as a *cohort sequential* design. To track adolescent brain, cognitive, and emotional development, the NCANDA study employed this design by initially recruiting youth in three age bands (12–14 years, 15–17 years, and 18–21 years) for subsequent annual testing [26]. The cognitive and motor functions assessed include executive functions, component processes of memory, social cognition, psychomotor speed, and visuospatial skills. Because the mainstay of the test battery is based on the computer-driven Web CNB (Computerized Neuropsychological Battery, [48, 49]), performance profiles of most of these component processes are measured in terms of accuracy, speed, and processing efficiency (the sum of standardized accuracy and speed scores) [50]. Thus, this battery assesses multiple cognitive domains, each of which can be subject to selective practice effects [2, 51].

Application of the twice-minus-once-tested method requires measurement of cross-sectional performance, for which the initial testing in NCANDA spanned ages 12 to 21 years [29], and longitudinal performance, which was measured annually using the same procedures and test materials [9]. The cross-sectional performance

provided the expected developmental effect free of prior experience over a decade of adolescent growth, and the longitudinal performance included the developmental effect plus learning. The difference between the second test score of an individual (twice tested) and the first test scores of the group at the individual's second test age (once tested) yielded an index of learning. This method revealed different extents of practice effects for the various test composites (29% to 99% of the variance was due to prior testing) and for different ages, where the younger participants showed the greatest improvement with little contribution from sex, ethnicity, or parental education [9].

The current analysis expanded the twice-minus-once-tested method to track development independent of learning from prior test experience in the NCANDA cohort over the first four years of the study. Recently, we used this approach to discriminate learning from development in the Stroop Match-to-Sample test [52], which assesses attentional inhibition, a function considered to advance over adolescence. Results indicated that learning contributed a greater proportion of the change variance than did development, which accounted for learning based prior testing [53]. The current analysis examined multiple component cognitive (Abstraction, Attention, Emotion, Episodic memory, Working memory, and General Ability) and motor (Ataxia and Speed) functions over the first four annual NCANDA test sessions. Accordingly, the first set of analyses quantified the effects of *learning* (i.e., prior test experience) on accuracy domain scores, pursuing three aims: 1) given previous longitudinal findings indicating significant learning (higher accuracy scores with prior experience) from initial repeated testing even with a year interval, we tested whether the amount and trajectory of learning between the second and later tests differed from those observed between the initial test pairs; 2) we questioned whether these parameters differed by functional domain; and 3) age and sex were examined as moderating factors. The same three aims were also applied to the speed measures with the expectation that improvement would be in the direction of faster response times; to put all accuracy and speed measures in the same direction, response speed was inverted so that larger values indicated faster performance. After quantifying learning effects, *developmental* effects were then determined for each domain for accuracy and speed, having removed the measured learning effects. We tested the hypothesis that the trajectories of scores would be different depending on the inclusion or removal of estimated practice effects, and that these trajectory differences would be present for accuracy and speed scores.

Methods

Participants

All participants were drawn from the NCANDA cohort of 692 who endorsed no or low levels of drinking (no-to-low alcohol drinkers) at baseline. The current longitudinal analysis required that each participant had 4 consecutive annual test sessions, starting from baseline and remained a no-to-low drinker (described below) for all included sessions. The resulting sample comprised 319 participants (154 male, 165 female), although not all participants had all composites; demographic descriptions are presented in Table 1.

All participants underwent informed consent processes at each visit with a research associate trained in human subject research protocols. Adult participants or the parents of minor participants provided written informed consent before starting the study; minor participants provided assent. The Institutional Review Boards of each site approved this study, and all methods were performed in accordance with the relevant guideline and regulations noted and approved.

Alcohol history and testing

Participants completed the Customary Drinking and Drug use Record (CDDR, [54]) to characterize past and current alcohol and substance use. At each test session, alcohol and drug use reports were accompanied by 12-panel urine toxicology screens 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 were sent for gas chromatography/mass spectrometry confirmation; if confirmed, participants were excluded from testing that day and from the current analysis.

To be considered a no-to-low drinker, participants met two sets of criteria determined with the CDDR described previously [55] as follows: 1) The *maximum lifetime drinking days* for male and female participants was ≤ 5 for age 12 to 15.9 years, ≤ 11 for age 16 to 16.9 years, ≤ 23 for age 17 to 17.9 years, and ≤ 51 for age 18 years old and older; and 2) The *maximum allowable drinks per occasion* was ≤ 3 for female participants at any age but varied by age for male participants: ≤ 3 for age 12 to 13.9 years, ≤ 4 for age 14 to 19.9 years, and ≤ 5 for age 20 years old and older.

Cognitive and motor tests and composite score construction

Assessment was the same across all five sites and used a combination of computerized tests (originally the Web CNB, now the WebCNP (https://webcnp.med.

Table 1 NCANDA demographics at baseline

Age (years)		
Male	**mean=**	15.0
	SD=	2.33
	Range=	12.0 to 21.2
	N=	154
Female	**mean=**	14.8
	SD=	2.21
	Range=	12.0 to 21.3
	N=	165
Socioeconomic status†	**mean=**	16.6
	SD=	2.51
Self-declared Ethnicity		
Caucasian	**N=**	230
African-American	**N=**	50
Asian	**N=**	34
Other	**N=**	5
Site		
UPitt	**N=**	51
SRI	**N=**	44
Duke	**N=**	62
OHSU	**N=**	70
UCSD	**N=**	92

†Highest education of a parent

upenn.edu/) [37, 48]) and traditional neuropsychological tests [29]. 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 Resource. 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 [56, 57] at SRI International. The longitudinal data used herein were available through a formal, locked data release (NCANDA_PUBLIC_3Y_REDCAP_V02).

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) [48]. Descriptions of the 15 WebCNP tests used were provided in our earlier report [29] (Supplemental Table 3 in Sullivan et al. 2016), with most tests having both accuracy and speed (response time) measures. A subset of measures from these tests was used to create theoretically-driven composite Z-scores for 8 accuracy measures (Abstraction, Attention, Emotion, Episodic Memory, Working Memory, General Ability, Balance, and Total) and 8 speed measures (Abstraction, Attention, Emotion, Episodic Memory, Working Memory, General Ability, Motor, and Total). In addition, an Efficiency score was calculated as the sum of the Total Accuracy plus Speed Z-scores [50]. The individual tests and computed composites were described previously, where Table 2 lists the cognitive and motor domains and specific processes assessed, with associated brain regions reported to support each process (see Supplemental table 2 in Sullivan et al. [29] also 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 [37, 58]. First, each measure was standardized on baseline scores achieved by all male and female adolescents who met NCANDA entry criteria (maximum $N=319$) and expressed as a Z-score (mean $=0\pm1$SD). 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; the number of participants

Table 2 R2 for each test session gamm and difference between test session pairs tested with ANOVA indicative of learning

		Time 1 to 2				
	N	Variance explained by	Variance explained by	% due to		
Composite Score	Male, Female	age + learning	learning	learning	L ratio	p-value
Accuracy						
Abstraction	153, 163	0.0563	0.0216	38.45	30.7637	**0.0001**
Attention	151, 164	0.1286	0.0239	18.60	24.6525	**0.0001**
Emotion	153, 163	0.0584	0.0084	14.40	15.7658	**0.0001**
Episodic Memory	154, 165	0.0452	0.0641	100.00	69.6839	**0.0001**
Working Memory	151, 163	0.0152	0.0164	100.00	15.5467	**0.0001**
General Ability	152, 164	0.1167	0.0162	13.89	14.8150	**0.0001**
Balance	137, 155	0.0383	—	—	1.1385	0.2860
Average	145, 156	0.1413	0.0835	59.08	87.2525	**0.0001**
Speed						
Abstraction	153, 163	0.0195	0.0180	92.04	20.2487	**0.0001**
Attention	151, 164	0.0891	0.0113	12.63	9.8839	**0.0017**
Emotion	153, 163	0.0193	0.0136	70.30	15.6171	**0.0001**
Episodic Memory	154, 165	0.0501	0.0349	69.63	40.0072	**0.0001**
Working Memory	151, 163	-0.0015	—	—	0.8168	0.3661
General Ability	152, 164	0.0157	0.0065	41.15	8.0878	**0.0045**
Motor	152, 164	0.1599	0.0671	41.96	81.5878	**0.0001**
Average	149, 159	0.0364	0.0038	10.41	15.6193	**0.0001**
Average Efficiency	145, 155	0.1362	0.0582	42.75	75.4346	**0.0001**
		Time 2 to 3				
	N	Variance explained by	Variance explained	% due to		
Composite Score	Male, Female	age + learning	by learning	learning	L ratio	p-value
Accuracy						
Abstraction	153, 163	0.0192	0.0029	15.06	12.4709	**0.0004**
Attention	151, 164	0.0320	0.0006	2.00	8.0822	**0.0045**
Emotion	153, 163	0.0186	0.0020	10.72	11.3899	**0.0001**
Episodic Memory	154, 165	0.0036	0.0119	100.00	13.8872	**0.0002**
Working Memory	151, 163	-0.0009	—	—	1.2620	0.2613
General Ability	152, 164	0.0812	—	—	4.1676	0.0412
Balance	137, 155	0.0146	—	—	0.0060	0.9382
Average	145, 156	0.0462	0.0197	42.59	26.6596	**0.0001**
Speed						
Abstraction	153, 163	0.0261	0.0268	100.00	30.3700	**0.0001**
Attention	151, 164	0.0447	—	—	6.8150	0.0090
Emotion	153, 163	0.0272	0.0245	90.34	27.5515	**0.0001**
Episodic Memory	154, 165	0.0383	0.0290	75.78	29.8533	**0.0001**
Working Memory	151, 163	0.0013	—	—	0.4863	0.4856
General Ability	152, 164	0.0181	—	—	6.2356	0.0125
Motor	152, 164	0.0971	0.0394	40.61	43.9287	**0.0001**
Average	149, 159	0.0441	0.0038	8.51	9.1862	**0.0024**
Average Efficiency	145, 155	0.0777	0.0232	29.81	28.6264	**0.0001**
		Time 3 to 4				
	N	Variance explained by	Variance explained	% due to		
Composite Score	Male, Female	age + learning	by learning	learning	L ratio	p-value
Accuracy						
Abstraction	153, 163	0.0031	0.0049	100.00	9.4767	**0.0021**
Attention	151, 164	0.0171	—	—	1.7143	0.1904
Emotion	153, 163	0.0005	—	—	4.7172	0.0299
Episodic Memory	154, 165	-0.0010	—	—	2.8918	0.0890
Working Memory	151, 163	-0.0014	—	—	1.4571	0.2274
General Ability	152, 164	0.0502	—	—	6.3269	0.0119
Balance	137, 155	-0.0002	—	—	2.9040	0.0884
Average	145, 156	0.0116	—	—	6.2556	0.0169
Speed						
Abstraction	153, 163	0.0055	—	—	0.3224	0.5702
Attention	151, 164	0.0179	—	—	4.5412	0.0331
Emotion	153, 163	-0.0002	—	—	1.1173	0.2905
Episodic Memory	154, 165	0.0041	—	—	3.5691	0.0589
Working Memory	151, 163	-0.0003	—	—	1.0582	0.3036
General Ability	152, 164	0.0036	—	—	1.0618	0.3028
Motor	152, 164	0.0428	—	—	4.5008	0.0339
Average	149, 159	0.0064	—	—	0.6613	0.4161
Average Efficiency	145, 155	0.0144	—	—	0.4709	0.4926

†Improvement in R2 between a pair of test sessions; see red values in Fig. 1 bar plots

Bold values are significant with a family-wise Bonferroni correction for 8 comparisons (alpha = 0.05) at $p \leq 0.00625$

% due to learning values are noted only for significant improvement

with scores are in Table 2 in the Results. 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; missing scores were allowed, but each composite score had to have at least 2 measures to make a domain.

Statistical analysis

The primary analysis tools were the General Additive Mixed Model (GAMM) and Likelihood Ratio Tests (LRT) using the *gamm* and *anova* functions from the *mgcv* package in R Version 3.1.0 [http://www.r-project.org]. Age was allowed to be a nonlinear smooth effect, implemented via thin plate splines [s(age)] with 3 internal knots [59], herein after referred to as "smoothed age." Roughness penalties for the smooth effects were estimated using generalized cross validation [60].

Estimation of learning from visit to visit

To determine the learning effect from visit to visit, three data sets were constructed: 1) all subjects' visit 1 plus visit 2; 2) all subjects' visit 2 plus visit 3; 3) all subjects' visit 3 plus visit 4. For each dataset, a regression analysis fitting the data with age (GAMM with smoothed age) was performed without and with visit as a factor. The results of the two models were compared with an LRT; significant improvement by adding visit to the model indicated significant learning from visit to the subsequent visit. The improvement in the amount of variance explained (R-squared) is reported here as an index of the amount of learning between visits.

To test for age effects on learning, for each test visit pair, GAMMs with and without an age-by-visit interaction were compared using LRTs. A significantly better fit with age-by-visit interaction indicated significant age effects on learning.

To test for sex effects on learning, for each visit pair, GAMMs with age-by-visit plus sex-by-visit were performed and examined for learning-by-sex interactions.

Learning-adjusted developmental model

To quantify learning across the four visits, a sequence of model fits was performed [53] that allowed the estimation of development effects independent of learning effects. The learning-adjusted development estimate across sessions was calculated as follows:

The *cross-sectional fit* of dependent variable y vs. age was computed across all participants for each visit separately producing: *fit1 (based on only 1st visits), fit2 (based on only 2nd visits), fit3 (based on only 3rd visits), fit4 (based on only 4th visits).*

For visits 2, 3, and 4, the *age-related learning effect* from the previous visit was estimated by computing the difference between the predicted values from the cross-sectional fit at the current visit minus the predicted value when applying the fit from the previous visit to the ages at the current visit. This procedure was done cumulatively across visits 2, 3, and 4, producing learning-adjusted (i.e., learning-removed) values. Because visit 1 had no learning relevant to these test sessions, visit 1 values were not adjusted for experience effects. This adjustment is a direct extension of the "once vs. twice tested" method to more than two testing occasions.

The estimated age-dependent learning at visit 2 was the difference between the predicted values of cross-sectional ***fit2*** applied to subject visit 2 ages minus the predicted values of cross-sectional ***fit1*** applied to the same visit 2 ages:

$$visit2.adj = visit2 - (predict(fit2\_on\_visit2) - predict(fit1\_on\_visit2))$$

This is the simple case of baseline with one follow-up test session as used in Sullivan et al. [9] and is the "once minus twice tested" method [6].

For subsequent visits the learning effect required testing of additional learning from visit to visit, calculated as follows: The estimate of learning at visit 3 was the difference between the predicted values of cross-sectional ***fit3*** applied to subject ages at visit 3 minus the predicted values of cross-sectional ***fit2*** again applied to subject ages at visit 3. This quantity was then added to the estimate from visit 1 to visit 2 to obtain the cumulative learning effect from baseline to visit 3:

$$\begin{aligned} visit3.adj = visit3 - [&(predict(fit2\_on\_visit2) - predict(fit1\_on\_visit2)) \\ &+ \\ &(predict(fit3\_on\_visit3) - predict(fit2\_on\_visit3))] \end{aligned}$$

The estimate of cumulative learning at visit 4 was the difference between the predicted values of cross-sectional ***fit4*** applied to subject ages at visit 4 minus the predicted values of cross-sectional ***fit3*** applied to the same subject ages at visit 4, which was then added to the estimate from visit 1 to visit 2 and from visit 2 to visit 3:

$$\begin{aligned} visit4.adj = visit4 - [&(predict(fit2\_on\_visit2) - predict(fit1\_on\_visit2)) \\ &+ \\ &(predict(fit3\_on\_visit3) - predict(fit2\_on\_visit3)) \\ &+ \\ &(predict(fit4\_on\_visit4) - predict(fit3\_on\_visit4))] \end{aligned}$$

To examine the effect of age on performance, GAMMs examining composite scores as a function of smoothed age were performed before and after adjusting for

learning. ANOVAs were computed to allow comparison of the GAMM with smoothed age to the GAMM with linear age to determine whether the developmental trajectory of that test composite was better describes as smoothed or linear.

To test for sex effects on development across all visits, GAMMs examining learning-adjusted values as a function of smoothed and linear age with and without sex as a factor were compared with ANOVA.

To account for the multiple comparisons made, family-wise Bonferroni correction was determined for 8 test session pairs for each metric (accuracy and speed) with $\alpha = 0.05$ required p-values ≤ 0.00625 (two-tailed) to be considered significant.

Results

The first set of results quantifies the learning effects for each performance metric (accuracy and speed) of each composite score by age and sex. The second set quantifies the developmental effects with the measured learning (i.e., practice) effects removed.

Learning effects

For each composite score, learning was quantified by computing the difference in the variance explained by age between pairs of two GAMM models with and without visit as a factor; these statistics are presented in Table 2 along with the percent change (typically improvement), their associated LRT L ratios of the model fits, and p-values. The additional variance explained by age plus visit in the model is indicative of learning and is depicted in red in the second bar of each visit pair in Fig. 1. The learning effect did not differ significantly by sex between test session pairs for any accuracy composite score but showed a modest sex effect for the Motor speed composite, described below (Table 3). The trend was for the younger participants to show greater learning than the older ones especially between sessions 1 and 2 (Figs. 2 and 3, column 5). Cross-sectional scores at each test session *over age* are presented in Figs. 2 and 3 (first of the 3 right panels), and the learning component is presented in the second and third figures of the right triplet for each composite score over age.

Accuracy composite scores

Overall, smooth age fits were better than linear age fits in describing the unadjusted data, which comprised both learning and development (Table 4; Fig. 2a-c, left spaghetti plots for each composite). Two exceptions were Working Memory and Balance; the latter showed a smooth trend.

For *Abstraction* accuracy, the ANOVA comparing the GAMM fits between each successive pairs of test sessions indicated significant performance improvement between each pair (red area of Fig. 1 and Table 2). In all three cases, significant age-by-learning interactions indicated that learning was greater with younger age.

For *Attention, Emotion*, *Episodic Memory* and *Average* accuracy, the ANOVA comparing the GAMM fit pairs of test sessions indicated significant improvement from time 1 to 2 and 2 to 3 but not from 3 to 4. Further, learning interacted with age, indicating greater learning with younger age. For *Working Memory* and *General Ability* accuracy, the ANOVA revealed significant improvement that interacted with age from time 1 to 2. *Balance* was the only composite failing to show significant improvement between any test pairs and no interaction with age.

Speed composite scores

Like Accuracy, Speed showed improvement over test sessions, but the overall pattern of improvement in composite scores differed by metric. Smooth age fits were better than linear age fits in describing the unadjusted data, which comprised both learning and development (Table 4; Fig. 3a-d, left spaghetti plots for each composite), for four of the eight test composites and the Efficiency score: Attention, Episodic Memory, Motor, and Average Speed. Linear fits better described the age effect for the four remaining composites: Abstraction, Emotion, Working Memory, and General Ability; Emotion and General Ability showed smooth trends.

The ANOVAs comparing the GAMM fits revealed significant increases in speeded responses between the first two pairs of test sessions but not the last pair for five composite scores and for Average Efficiency: Abstraction, Emotion, Episodic Memory, Motor, and Average Speed. Attention and General Ability speed improved from time 1–2 only. Working Memory showed no improvement between any test pair (Fig. 3a-d, 3 plots in the right panel; Table 2).

The age-learning interaction was significant for time 1–2 and time 2–3 for Abstraction, Emotion, Episodic Memory, Motor, and Average Speed (Table 3, Fig. 3a-d). For General Ability speed, the interaction with age was significant between time 1–2 and showed a trend between time 2–3.

The learning effect in the speed scores between test session pairs differed by sex for the Motor composite only. The sex difference occurred between tests 1 and 2 and indicated that the female participants showed a greater gain in speed than the male participants (see differences in confidence intervals for female scores in red relative to male scores in blue in Fig. 3, left panel).

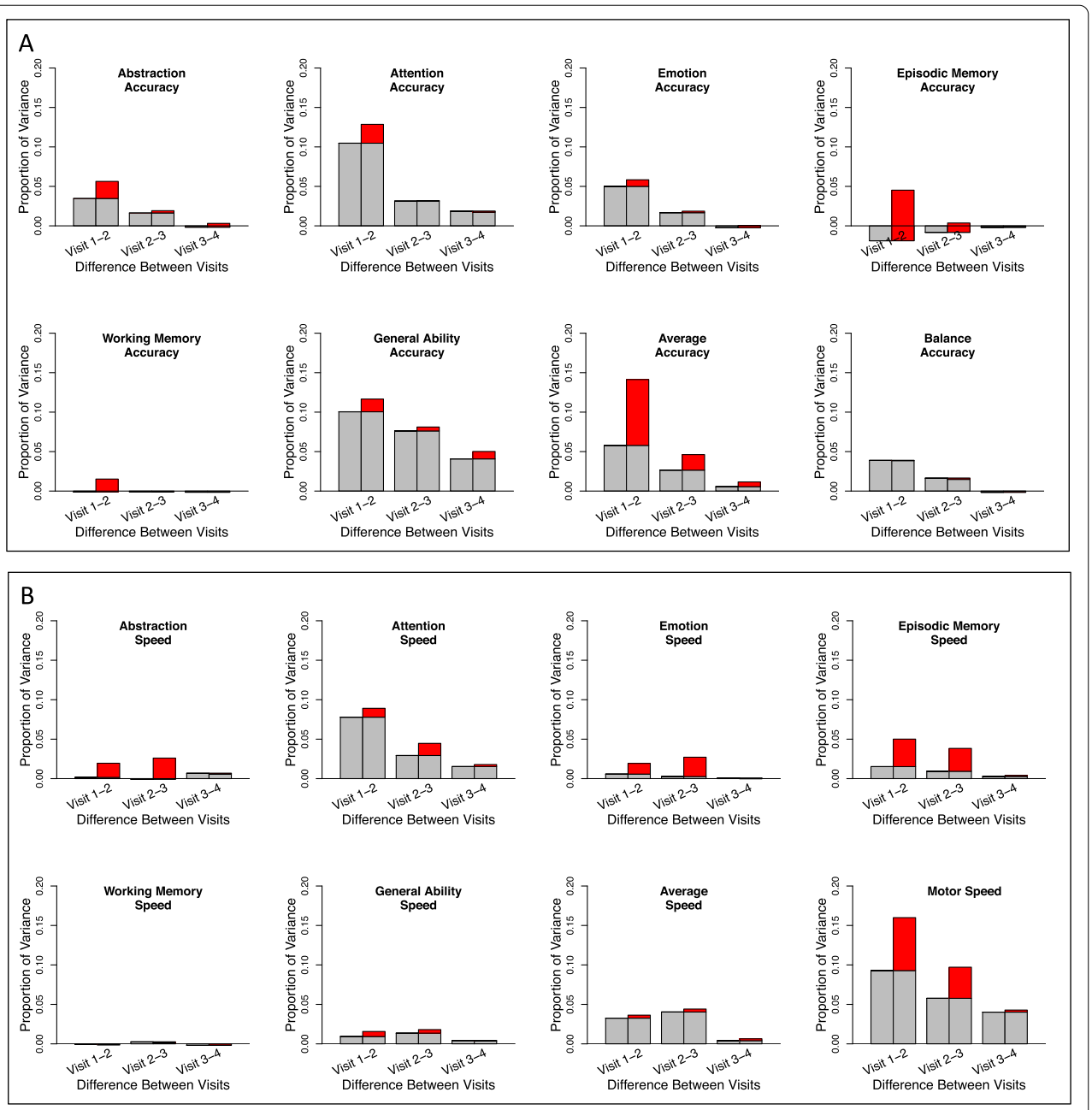

Fig. 1 The difference in variance explained by age between each visit pair (e.g., visit 1 compared to visit 2) with age alone in the left bar of each pair and age + visit in the right bar of each pair, with the additional variance explained by learning depicted in red in the right bar of each visit pair

(See figure on next page.)

Fig. 2 a-c Two left panels: The gray spaghetti plots show accuracy performance of each person for each of the four test sessions for each test composite. The gray regression lines indicate the ± 1 and ± 2 standard deviations of all participants. The color regression lines indicate the mean and 95% confidence interval of the performance by male (blue) and female (red) participants. The left plots show the learning + developmental effect; the right plots show the learning-adjusted developmental effect. Three right panels depict learning by session in accuracy scores. The first plot presents the fit of the cross-sectional scores at each test session over age: black = test 1, red = test 2, green = test3, and blue = test 4. The second plot displays the learning between tests 1–2 (red), tests 2–3 (green), and tests 3–4 (blue) over age. The third plot also displays the learning over age between test pairs normalized at 0 to reveal age effects and their differences between test pairs. The general trend was for the younger participants to show greater learning than the older ones especially between sessions 1 and 2 (red filled plots)

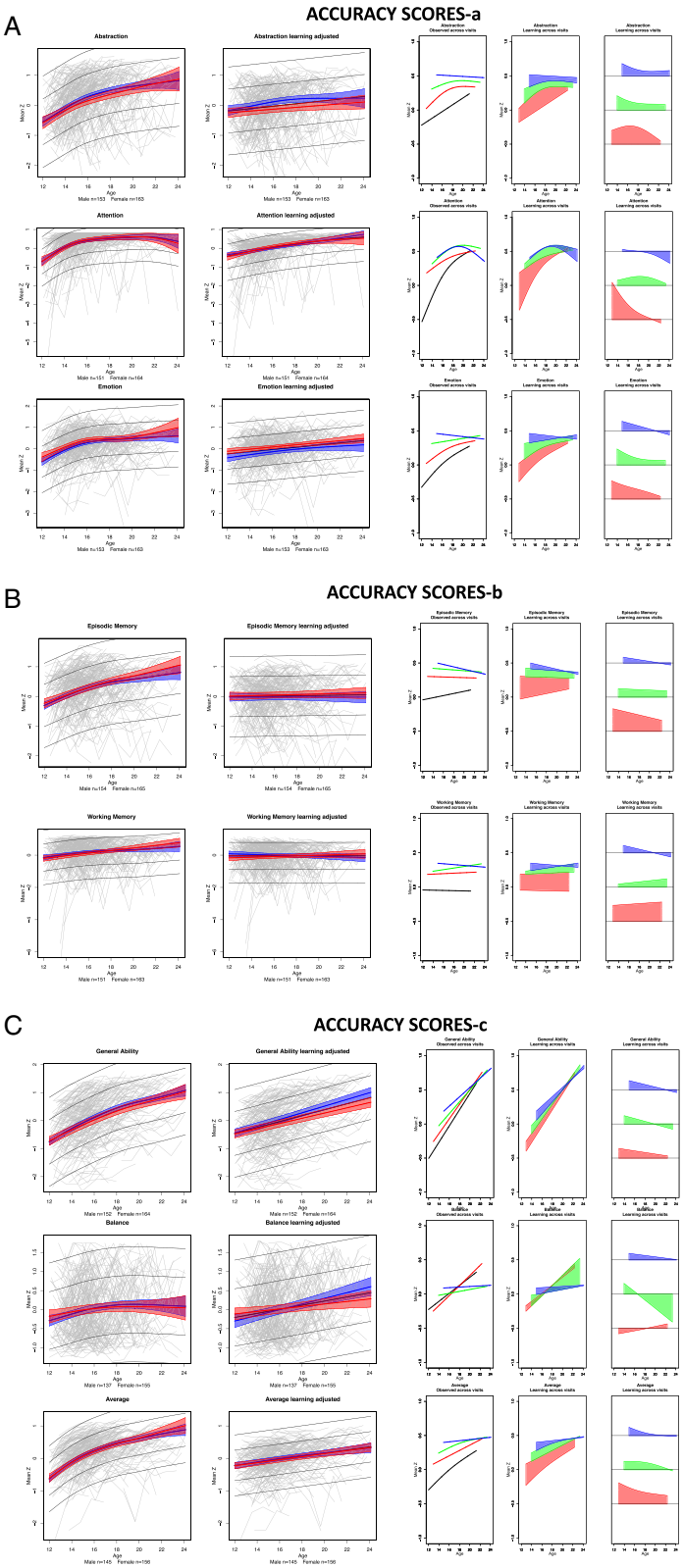

Fig. 2 (See legend on previous page.)

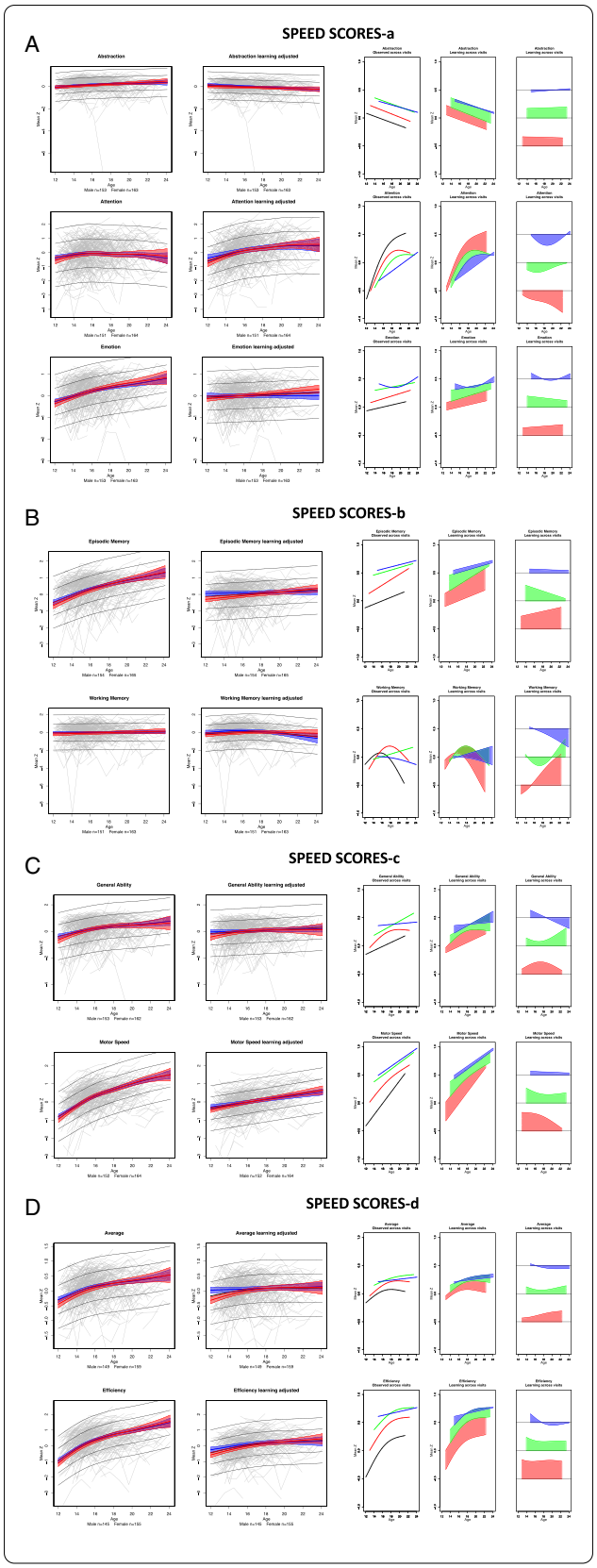

Fig. 3 **a-d** Two left panels: The gray spaghetti plots show speed performance of each person for each of the four test sessions for each test composite. The gray regression lines indicate the ± 1 and ± 2 standard deviations of all participants. The color regression lines indicate the mean and 95% confidence interval of the performance by male (blue) and female (red) participants. The left plots show the learning + developmental effect; the right plots show the learning-adjusted developmental effect. Three right panels depict learning by session in speed scores. The first plot presents the fit of the cross-sectional scores at each test session over age: black = test 1, red = test 2, green = test3, and blue = test 4. The second plot displays the learning between tests 1–2 (red), tests 2–3 (green), and tests 3–4 (blue) over age. The third plot also displays the learning over age between test pairs normalized at 0 to reveal age effects and their differences between test pairs. Unlike the accuracy scores, the general trend for the speed scores showed different age trends for the different test composites

Learning-adjusted developmental effects

Accuracy composite scores

These results describe developmental effects for each composite after removing the estimated learning effects. For all 8 composite accuracy scores, the developmental effect was significant across age and was adequately described by linear fits with no improvement from nonlinear (smooth) fits ($p = 0.999$ for all composites except Attention $p = 0.173$) (Table 4; Fig. 2, right panel of spaghetti plots). Thus, the accelerated improvement in scores over age (Fig. 2 spaghetti plots in left panels) was attributed to greater learning rather than apparent accelerated development in the younger relative to the older participants.

Speed composite scores

The learning-adjusted developmental effects for speed were adequately described by linear fits with no improvement from smooth fits ($p = 0.999$) for *Abstraction, Emotion, Episodic Memory, General Ability,* and *Motor* scores. The smooth fit was better than the linear fit for *Attention, Working Memory, Average Speed*, and *Efficiency* (Table 4; Fig. 3 right spaghetti plots).

The effect of learning adjustment can also be portrayed by comparing the cross-sectional age relation to the longitudinal age relation with and without learning adjustment as per Salthouse [61]. Figure 4 presents the average slope from the simple cross-sectional linear regression at baseline (value in Z units/year) compared to the fixed effects from a linear mixed-model regression of the data across all 4 years before and after learning adjustment for the accuracy and speed domains. With few exceptions (notably, Attention speed; but even in this instance, the learning-adjusted better reflected the cross-sectional results than ignoring the practice), the non-adjusted data overestimated the rate of change per year and the learning adjusted more closely reflected the initial cross-section age relation.

Table 3 Interactions of age or sex with learning between test session pairs

	Age-by-learning interaction						Age-by-learning-by-sex interaction					
	Test session 1-2		Test session 2-3		Test session 3-4		Test session 1-2		Test session 2-3		Test session 3-4	
Composite Score	L ratio	p-value	L ratio	p-value	L ratio	p-value	t value	p-value	t value	p-value	t value	p-value
Accuracy												
Abstraction	30.8219	**0.0001**	12.7195	**0.0004**	8.6790	**0.0032**	0.1960	0.8447	1.5208	0.1288	0.3397	0.7342
Attention	26.8854	**0.0001**	8.3190	**0.0039**	1.7143	0.1904	-0.0167	0.9867	-1.1137	0.2658	1.5148	0.1303
Emotion	17.3454	**0.0001**	10.9633	**0.0009**	2.9997	0.0833	0.2114	0.8326	1.1720	0.2416	-1.0386	0.2994
Episodic Memory	52.7130	**0.0001**	11.6998	**0.0006**	1.7360	0.1876	0.2030	0.8392	-0.0832	0.9337	0.1004	0.9201
Working Memory	15.5141	**0.0001**	1.1539	0.2827	0.8004	0.3710	-1.0668	0.2676	-1.1095	0.2676	0.6835	0.4945
General Ability	13.9813	**0.0002**	1.7997	0.1797	6.5742	0.0103	0.9461	0.3444	0.6901	0.4904	0.7027	0.4825
Balance	0.1802	0.6712	13.6364	**0.0002**	2.5859	0.1078	-1.0415	0.2981	1.8957	0.0585	1.0396	0.2990
Average	56.7172	**0.0001**	28.5127	**0.0001**	5.7068	0.0169	-0.0859	0.9315	-0.2187	0.8270	1.0281	0.3043
Speed												
Abstraction	20.5949	**0.0001**	30.3287	**0.0001**	0.1953	0.6585	-0.1014	0.9193	0.7039	0.4817	0.4243	0.6715
Attention	5.5566	0.0184	5.9795	0.0145	4.5412	0.0675	-0.2673	0.7893	-2.3194	0.0207	1.1935	0.2331
Emotion	15.4371	**0.0001**	27.1539	**0.0001**	0.0758	0.7831	-1.8774	0.0609	-0.9352	0.3500	0.2054	0.8373
Episodic Memory	39.0001	**0.0001**	18.2520	**0.0001**	3.5215	0.0606	-0.2993	0.7648	-1.5020	0.1336	-0.4478	0.6545
Working Memory	3.2621	0.0709	0.0926	0.7608	2.6763	0.1019	-0.9789	0.3280	0.2170	0.8283	-1.2041	0.2290
General Ability	8.0007	**0.0047**	6.2347	0.0125	3.1613	0.0754	1.3724	0.1704	-1.9202	0.0553	-0.4062	0.6848
Motor	55.2280	**0.0001**	38.9515	**0.0001**	4.4046	0.0358	-2.8843	**0.0041**	0.0355	0.9717	1.3220	0.1867
Average	16.6667	**0.0001**	15.9750	**0.0001**	2.7253	0.0988	-0.4519	0.6515	-2.0444	0.0413	-0.0038	0.9970
Average Efficiency	49.8311	**0.0001**	33.7423	**0.0001**	2.1940	0.1385	-0.7056	0.4807	-1.3712	0.1708	0.3758	0.7072

NB: See left panel of spaghetti plots for learning + development

Bold values are significant with a family-wise Bonferroni correction for 8 comparisons (alpha = 0.05) at t $p \leq 0.00625$

Table 4 Test for linear vs. smooth fit across all sessions (with sex in the model) for development with and without learning effects

	Learning + development				Learning-adjusted development		
Composite Speed Score	Curve fit	L ratio	p-value	Curve fit	Linear slope†	L ratio	p-value
Accuracy							
Abstraction	smooth	28.9024	**0.0001**	linear	0.0397	1.13E-06	0.9992
Attention	smooth	73.5215	**0.0001**	linear	0.0912	1.85675	0.1730
Emotion	smooth	40.9218	**0.0001**	linear	0.0519	1.22E-06	0.9991
Episodic Memory	smooth	16.5901	**0.0001**	linear	0.0054	6.81E-07	0.9993
Working Memory	linear	0.6166	0.4323	linear	-0.0001	1.07E-06	0.9992
General Ability	smooth	41.9402	**0.0001**	linear	0.1059	9.14E-07	0.9992
Balance	linear	7.1606	0.0075	linear	0.0538	9.13E-07	0.9992
Average	smooth	91.6728	**0.0001**	linear	0.0475	6.06E-07	0.9994
Speed							
Abstraction	linear	1.0527	0.3049	linear	-0.0280	8.36E-07	0.9993
Attention	smooth	9.5562	**0.0020**	smooth	0.1162	14.84196	**0.0001**
Emotion	linear	4.1107	0.0426	linear	0.0185	1.03E-06	0.9992
Episodic Memory	smooth	8.6477	**0.0033**	linear	0.0325	1.24E-06	0.9991
Working Memory	linear	0.0000	0.9991	smooth	0.0125	13.4889	**0.0000**
General Ability	linear	4.9148	0.0266	linear	0.0381	1.20E-06	0.9991
Motor	smooth	49.8218	**0.0001**	linear	0.0819	6.65E-07	0.9993
Average	smooth	19.1959	**0.0001**	smooth	0.0290	8.02164	**0.0046**
Average Efficiency	smooth	87.3808	**0.0001**	smooth	0.0785	15.30697	**0.0001**

Bold values are significant with a family-wise Bonferroni correction for 8 comparisons (alpha = 0.05) at $p \leq 0.00625$

See Figs. 2 and 3 spaghetti plots: left panel = learning + development; right plots = learning-adjusted development

†Slopes are taken from linear models and estimate the Z-unit change per year

Discussion

In general, mean scores for both accuracy and speed improved on most composite scores across the four test sessions with the greatest gains between the first and second tests (Fig. 2, two right panels). Abstraction was the only composite to show improvement over all 3 comparisons. Except for Working Memory, the overall learning for the Accuracy composite scores was greater in younger than older participants (Fig. 2, column 5). The Balance composite, which assessed gait and quiet standing, was unique in showing no significant change between any test session. Learning was detectable even in the context of apparent ceiling effects in the Attention and Working Memory composites (Fig. 2, spaghetti plots). Unlike the Accuracy scores, the Speed composite scores did not show greater learning in the younger than older participants, except for the Motor Speed composite.

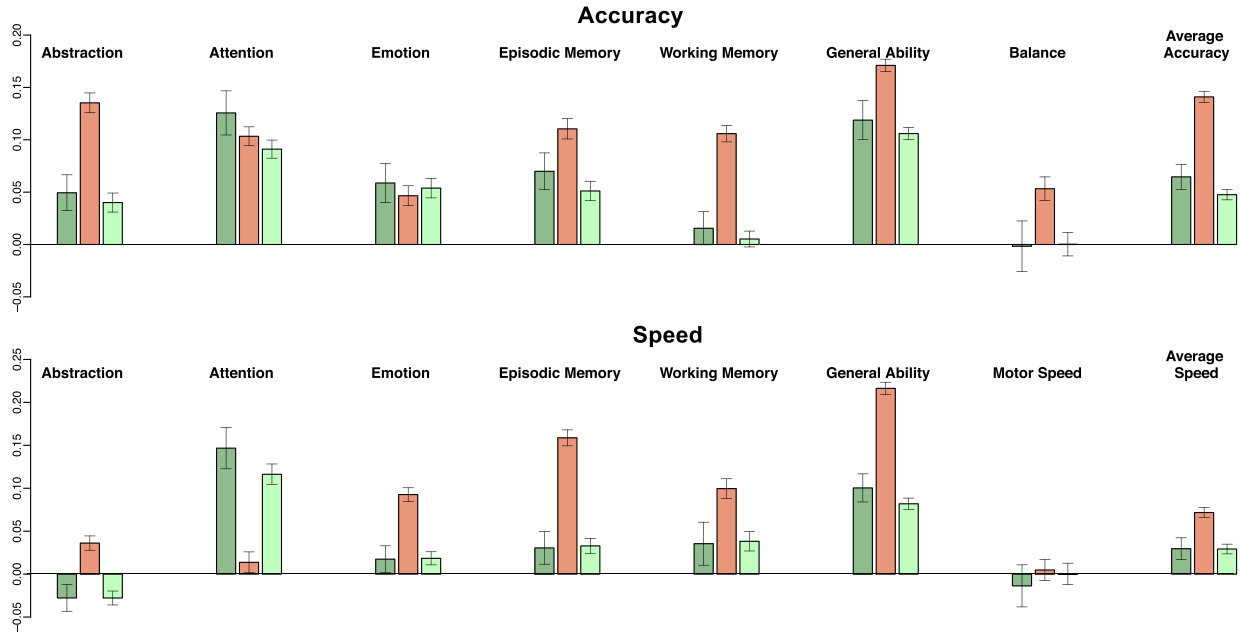

Fig. 4 Top: Accuracy; bottom: Speed. Dark green = rate of change/year from cross-sectional analysis of baseline data. Salmon = rate of change/year from fixed effects of mixed-model analysis of data across all 4 years. Light green = rate of change/year from fixed effects of mixed-model analysis of learning-adjusted data across all 4 years

We tested the hypothesis that the trajectories of scores would be different depending on the inclusion or removal of estimated practice effects, and that these trajectory differences would be present for accuracy and speed scores. For most composite scores, these hypotheses were supported. Specifically, developmental effects were linear for all accuracy composites. Thus, what appeared as accelerated advancement in the younger boys and girls was attributable to the learning component of the aging functions. As with the linear trends for accuracy, five learning-adjusted speed scores were linear. Speed scores showing smooth fits were Attention, Working Memory, Average Speed, and Efficiency; each exhibited a slight inflection, notably in the younger ages that were modestly steeper for the girls than the boys. In all cases, the confidence intervals of the female participants overlapped or exceeded the speed score intervals of their male counterparts, albeit non-significant from the GAMM analyses, thereby lending little support for a male advantage in speeded responding in cognitive or motor realms in later adolescence (cf. [27]). Further, sex differences may be attenuated with multiple annual test sessions, noted herein. Absence of age-by-sex interactions was also reported in a 4-year longitudinal study of youth tested every 2 years (ages 6 to 18 years at baseline testing) despite sex-related performance differences in specific tests: male youth achieved better scores than female youth on Block Design, whereas the opposite occurred on Grooved Pegboard and Digit Symbol Coding tests [11]. One interpretation is that the sex differences were stable despite repeated testing and presumed further development over the 4-year interval.

The original, cross-sectional analysis of the WebCNP composite scores noted that sex differences had smaller effect sizes than age but were evident, with female participants outperforming their male counterparts on attention, word and face memory, reasoning speed, and all social cognition tests, whereas male participants outperformed their female counterparts in spatial processing and sensorimotor and motor speed [37]. Comporting with those cross-sectional findings, our current longitudinal observations revealed that these sex differences were greatest at younger ages, with adolescent development, female participants became faster over time on Motor Speed.

To quantify practice effects associated with subtests of the computer-based test battery *Cognition,* which is based on the Web CNP, Basner and colleagues [15] varied testing parameters, including test forms and test–retest intervals for retesting upwards of 15 times. Remarkably, even their 6 subtests using unique stimuli in subsequent test sessions evidenced practice effects, consistent with the interpretation that some form of procedural learning beyond episodic memory for specific test information contributes to practice effects, that is, prior experience.

In the current study, the variance explained by age + learning in the visit-to-visit analyses ranged from less than 0 to 16%. Within those totals, the proportion of accuracy score variance attributable to *learning* ranged from 14% for Emotion to more than 100% for the two Memory composites between test sessions 1 and 2 (Fig. 1 and Table 2). For Average accuracy, the gain was nearly 60% from tests 1 to 2 and remained high between tests 2 to 3 at 42%. Learning-associated improvement in speed scores between sessions 1 and 2 was especially high for Abstraction (92%), Emotion (70%), and Episodic Memory (69%). Thus, a unique contribution of this analysis was to address whether practice effects in this adolescent to young adult age range would accrue beyond the first follow-up testing and, if so, would occur in all functional domains examined. Although the greatest learning effect occurred between the first and second visits, further learning was measurable between visits 2 and 3 and again between visits 3 and 4 for two accuracy scores (General Ability and Average) and for several speed scores and the Efficiency score, as depicted in the red segments in Fig. 1.

The slopes from the linear models describe the *learning-adjusted developmental* performance trajectories (Table 4) in terms of Z-unit changes per year (Figs. 2 and 3 right spaghetti plots). Extrapolating from the youngest to the oldest youth, the functional composites showing the largest improvement were General Ability accuracy and Attention accuracy and speed, with estimated gains of approximately 1.0 Z-unit over a decade. The developmental estimates would have been inflated had they not been adjusted for learning. Indeed, a benefit of the cohort sequential design in longitudinal assessment was demonstrated in our analysis, based on Salthouse [61], that compared cross-sectional slopes, longitudinal slopes unadjusted for practice effects, and longitudinal slopes adjusted for practice effects (Fig. 4). Use of the twice minus once tested analysis enabled this comparison, which revealed that for most composites the longitudinal slopes adjusted for practice effects reflected the cross-sectional slopes, a pattern previously noted in a longitudinal analysis of cognitive performance by men and women spanning the adult age range [1]. By contrast, the unadjusted longitudinal slopes were substantially greater and thus over-estimated the developmental trajectories. Critically, longitudinal sessions initiated at a single or narrow age preclude such an adjustment, which requires cross-sectional observations to be made over wide age bands.

Limitations

Although use of composite scores can reduce excessive variance often observed in individual tests, the test composites created in the current study, which were similar to those used by Gur and colleagues [48–50, 62], comprised different numbers of measures that may have contributed to differences in variances. Further, some tests may be more difficult than others, and difficulty levels may differ by variables such as age, sex, or individual abilities. Despite the strength of the twice-minus-once-tested method, representation of each age in adolescence had a limited sample size, which was then halved in the sex analyses. This method may also be subject to cohort differences by recruitment age bands [1, 2, 63].

Conclusion

Longitudinal study, held as the gold standard for tracking developmental trajectories, must take prior assessment experience, also considered learning or practice effects, into account. Study protocols that recruit all participants at one age or within a narrow age band are not positioned to use the twice-minus-once-tested method to dissociatate learning from development, whereas studies using the cohort sequential (accelerated longitudinal) design are poised to do so. Had our method not dissociated the effects of learning from development, the course of developmental changes over the adolescent years would have been interpreted as following an accelerating increase, notable in younger ages. By contrast, removing the learning effects revealed a linear developmental trajectory for all accuracy composite scores for all cognitive functions examined. Recognition of the substantial influence of prior testing experience, which does not necessarily rely on repetition and memory for specific test items (cf., [15]) and can be a metric of interest in its own right [8], is critical to be accomplished in highly vetted groups of adolescents and emerging adults. Doing so will enable accurate characterization of normal development and provide norms for other uses, including clinical neuropsychological investigations of conditions affecting the brain whatever the cause.

Abbreviations

NCANDA: National Consortium on Alcohol and NeuroDevelopment in Adolescence; CDDR: Customary Drinking and Drug use Record; WebCNB: Web Computerized Neurocognitive Battery; WebCNP: Web Computerized Neuropsychological Testing System; GAMM: General Additive Mixed Model; LRT: Likelihood Ratio Tests; ANOVA: Analysis of variance.

Acknowledgements

Not applicable.

Authors' contributions

Edith V. Sullivan, Ph.D., Wesley K. Thompson, Ph.D., Ty Brumback, Ph.D., Devin Prouty, Ph.D., Susan F. Tapert, Ph.D., Sandra A. Brown, Ph.D., Michael D. De Bellis, M.D., Kate B. Nooner, Ph.D., Fiona C. Baker, Ph.D., Ian M. Colrain, Ph.D., Duncan B. Clark, M.D., Ph.D., Bonnie J. Nagel, Ph.D., Kilian M. Pohl, Ph.D., Adolf Pfefferbaum, M.D. EVS, WKT (biostatistician), AP conducted the primary analyses and wrote and produced the first draft of the manuscript, tables, and figures. KMP curated the data and assembled the data release. TB, DP, SFT, SAB, MDDB, KBN, FCB, IMC, DBC, BJN participated in data collection and data transfer to

the Data Analysis Resource at SRI International. All authors participated in the design of the NCANDA study and read, were given the opportunity to edit the manuscript, and approved the final draft.

Funding
Data collection, release, analysis, and write up were supported by the U.S. National Institute of Health funding [AA021697 (KMP + AP), AA021695 (SFT + SAB), AA021692 (SFT), AA021696 (FCB + IMC), AA021681 (MDB), AA021690 (DBC), AA021691 (BN)].

Availability of data and materials
The data were part of the public data release NCANDA_PUBLIC_3Y_REDCAP_V02*, distributed according to the NCANDA Data Distribution agreement.**
* Pohl KM, Sullivan EV, Podhajsky S, Baker FC, Brown SA, Clark DB, Colrain IM, De Bellis MD, Nagel BJ, Nooner KB, Tapert SF, Pfefferbaum A: The `NCANDA_PUBLIC_3Y_REDCAP_V04` Data Release of the National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA), Sage Bionetworks Synapse. https://dx.doi.org/10.7303/syn26350358
**https://www.niaaa.nih.gov/ncanda-data-distribution-agreement

Declarations

Ethics approval and consent to participate
All participants underwent informed consent processes at each visit with a research associate trained in human subject research protocols. Adult participants or the parents of minor participants provided written informed consent before starting the study; minor participants provided assent. The Institutional Review Boards of each study site approved this study: UC San Diego Human Research Protections Program (HRPP); SRI International, Advarra, Inc. (FWA number 00023875); Duke University School of Medicine Institutional Review Board; University of Pittsburgh Institutional Review Board; Oregon Health & Sciences University Institutional Review Board.

Consent for publication
All authors consented to publication.

Competing interests
None of the authors declare any conflict of interest regarding the content of the manuscript.

Author details
[1]Department of Psychiatry & Behavioral Sciences, Stanford University School of Medicine (MC5723), 401 Quarry Road, Stanford, CA 94305-5723, USA. [2]Division of Biostatistics and Dept of Radiology, University of California, San Diego, La Jolla, CA, USA. [3]Department of Psychological Sciences, Northern Kentucky University, Highland Heights, KY, USA. [4]Center for Health Sciences, SRI International, Menlo Park, CA, USA. [5]Department of Psychiatry, University of California, San Diego, La Jolla, CA, USA. [6]Department of Psychiatry & Behavioral Sciences, Duke University School of Medicine, Durham, NC, USA. [7]Department of Psychology, University of North Carolina Wilmington, Wilmington, NC, USA. [8]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA. [9]Departments of Psychiatry and Behavioral Neuroscience, Oregon Health & Sciences University, Portland, OR, USA.

Received: 26 October 2021 Accepted: 14 April 2022
Published online: 24 June 2022

References

1. Salthouse TA. Trajectories of normal cognitive aging. Psychol Aging. 2019;34(1):17–24.
2. McArdle JJ, Ferrer-Caja E, Hamagami F, Woodcock RW. Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. Dev Psychol. 2002;38(1):115–42.
3. Pfefferbaum A, Rohlfing T, Rosenbloom MJ, Chu W, Colrain IM, Sullivan EV. Variation in longitudinal trajectories of regional brain volumes of healthy men and women (ages 10 to 85 years) measured with atlas-based parcellation of MRI. Neuroimage. 2013;65:176–93.
4. Zhang Y, Kwon D, Esmaeili-Firidouni P, Pfefferbaum A, Sullivan EV, Javitz H, Valcour V, Pohl KM. Extracting patterns of morphometry distinguishing HIV associated neurodegeneration from mild cognitive impairment via group cardinality constrained classification. Hum Brain Mapp. 2016;37(12):4523–38.
5. Salthouse TA. Effects of first occasion test experience on longitudinal cognitive change. Dev Psychol. 2013;49(11):2172–8.
6. Salthouse TA. Test experience effects in longitudinal comparisons of adult cognitive functioning. Dev Psychol. 2015;51(9):1262–70.
7. Salthouse TA, Tucker-Drob EM. Implications of short-term retest effects for the interpretation of longitudinal change. Neuropsychology. 2008;22(6):800–11.
8. Duff K, Callister C, Dennett K, Tometich D. Practice effects: a unique cognitive variable. Clin Neuropsychol. 2012;26(7):1117–27.
9. Sullivan EV, Brumback T, Tapert SF, Prouty D, Fama R, Thompson WK, Brown SA, Cummins K, Colrain IM, Baker FC, et al. 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. Dev Cogn Neurosci. 2017;24:72–83.
10. Sirois PA, Posner M, Stehbens JA, Loveland KA, Nichols S, Donfield SM, Bell TS, Hill SD, Amodei N, Hemophilia G, et al. 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. 2002;27(2):121–31.
11. Waber DP, Forbes PW, Almli CR, Blood EA. Brain Development Cooperative G: 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. 2012;18(2):179–90.
12. Anderson M, Reid C, Nelson J. Developmental changes in inspection time: what a difference a year makes. Intelligence. 2001;29:475–86.
13. Hsieh S, Yang MH. Two-Year Follow-Up Study of the Relationship Between Brain Structure and Cognitive Control Function Across the Adult Lifespan. Front Aging Neurosci. 2021;13:655050.
14. Calamia M, Markon K, Tranel D. Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. Clin Neuropsychol. 2012;26(4):543–70.
15. Basner M, Hermosillo E, Nasrini J, Saxena S, Dinges DF, Moore TM, Gur RC. Cognition test battery: Adjusting for practice and stimulus set effects for varying administration intervals in high performing individuals. J Clin Exp Neuropsychol. 2020;42(5):516–29.
16. Gavett BE, Ashendorf L, Gurnani AS. Reliable Change on Neuropsychological Tests in the Uniform Data Set. J Int Neuropsychol Soc. 2015;21(7):558–67.
17. Sawrie SM, Chelune GJ, Naugle RI, Luders HO. Empirical methods for assessing meaningful neuropsychological change following epilepsy surgery. J Int Neuropsychol Soc. 1996;2(6):556–64.
18. Elman JA, Jak AJ, Panizzon MS, Tu XM, Chen T, Reynolds CA, Gustavson DE, Franz CE, Hatton SN, Jacobson KC, et al. Underdiagnosis of mild cognitive impairment: A consequence of ignoring practice effects. Alzheimers Dement (Amst). 2018;10:372–81.
19. Estevis E, Basso MR, Combs D. Effects of practice on the Wechsler Adult Intelligence Scale-IV across 3- and 6-month intervals. Clin Neuropsychol. 2012;26(2):239–54.
20. Goldberg TE, Harvey PD, Wesnes KA, Snyder PJ, Schneider LS. Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials. Alzheimers Dement (Amst). 2015;1(1):103–11.
21. Naugle RI, Chelune GJ, Cheek R, Luders H, Awad IA. Detection of changes in material-specific memory following temporal lobectomy using the Wechsler Memory Scale-Revised. Arch Clin Neuropsychol. 1993;8(5):381–95.
22. Hedden T, Gabrieli JD. Insights into the ageing mind: a view from cognitive neuroscience. Nat Rev Neurosci. 2004;5(2):87–96.
23. Suchy Y, Kraybill ML, Franchow E. Practice effect and beyond: reaction to novelty as an independent predictor of cognitive decline among older adults. J Int Neuropsychol Soc. 2011;17(1):101–11.
24. Duff K, Lyketsos CG, Beglinger LJ, Chelune G, Moser DJ, Arndt S, Schultz SK, Paulsen JS, Petersen RC, McCaffrey RJ. Practice effects predict

cognitive outcome in amnestic mild cognitive impairment. Am J Geriatr Psychiatry. 2011;19(11):932–9.
25. Simmons C, Conley MI, Gee DG, Baskin-Sommers A, Barch DM, Hoffman EA, Huber RS, Iacono WG, Nagel BJ, Palmer CE, et al. Responsible Use of Open-Access Developmental Data: The Adolescent Brain Cognitive Development (ABCD) Study. Psychol Sci. 2021;32(6):866–70.
26. Brown SA, Brumback T, Tomlinson K, Cummins K, Thompson WK, Nagel BJ, De Bellis MD, Clark DB, Chung T, Hasler BP, et al. The National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA): A multi-site study of adolescent development and substance use. J Stud Alcohol Drugs. 2015;76(6):895–908.
27. Gur RC, Gur RE. Complementarity of sex differences in brain and behavior: From laterality to multimodal neuroimaging. J Neurosci Res. 2017;95(1–2):189–99.
28. Stiles J, Jernigan TL. The basics of brain development. Neuropsychol Rev. 2010;20(4):327–48.
29. Sullivan EV, Brumback T, Tapert SF, Fama R, Prouty D, Brown SA, Cummins K, Thompson WK, Colrain IM, Baker FC, et al. 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. 2016;30(4):449–73.
30. Akshoomoff N, Newman E, Thompson WK, McCabe C, Bloss CS, Chang L, Amaral DG, Casey BJ, Ernst TM, Frazier JA, et al. The NIH Toolbox Cognition Battery: results from a large normative developmental sample (PING). Neuropsychology. 2014;28(1):1–10.
31. Noble KG, Houston SM, Brito NH, Bartsch H, Kan E, Kuperman JM, Akshoomoff N, Amaral DG, Bloss CS, Libiger O, et al. Family income, parental education and brain structure in children and adolescents. Nat Neurosci. 2015;18(5):773–8.
32. Noble KG, Houston SM, Kan E, Sowell ER. Neural correlates of socioeconomic status in the developing human brain. Dev Sci. 2012;15(4):516–27.
33. Cole TJ, Pan H, Butler GE. A mixed effects model to estimate timing and intensity of pubertal growth from height and secondary sexual characteristics. Ann Hum Biol. 2014;41(1):76–83.
34. Tanner JM, Whitehouse RH, Takaishi M. Standards from birth to maturity for height, weight, height velocity, and weight velocity: British children, 1965. II Arch Dis Child. 1966;41(220):613–35.
35. Neligan G, Prudham D. Norms for four standard developmental milestones by sex, social class and place in family. Dev Med Child Neurol. 1969;11(4):413–22.
36. Hurks PP, Schrans D, Meijs C, Wassenberg R, Feron FJ, Jolles J. Developmental changes in semantic verbal fluency: analyses of word productivity as a function of time, clustering, and switching. Child Neuropsychol. 2010;16(4):366–87.
37. Gur RC, Richard J, Calkins ME, Chiavacci R, Hansen JA, Bilker WB, Loughead J, Connolly JJ, Qiu H, Mentch FD, et al. Age group and sex differences in performance on a computerized neurocognitive battery in children age 8–21. Neuropsychology. 2012;26(2):251–65.
38. Lawrence K, Campbell R, Skuse D. Age, gender, and puberty influence the development of facial emotion recognition. Front Psychol. 2015;6:761.
39. Piper BJ, Acevedo SF, Edwards KR, Curtiss AB, McGinnis GJ, Raber J. Age, sex, and handedness differentially contribute to neurospatial function on the Memory Island and Novel-Image Novel-Location tests. Physiol Behav. 2011;103(5):513–22.
40. Masters MS, Sanders B. Is the gender difference in mental rotation disappearing? Behav Genet. 1993;23(4):337–41.
41. Voyer D, Voyer S, Bryden MP. Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. Psychol Bull. 1995;117(2):250–70.
42. Denckla MB. Development of speed in repetitive and successive finger-movements in normal children. Dev Med Child Neurol. 1973;15(5):635–45.
43. Denckla MB. Development of motor co-ordination in normal children. Dev Med Child Neurol. 1974;16(6):729–41.
44. Piper BJ. Age, handedness, and sex contribute to fine motor behavior in children. J Neurosci Methods. 2011;195(1):88–91.
45. Dodds RM, Syddall HE, Cooper R, Benzeval M, Deary IJ, Dennison EM, Der G, Gale CR, Inskip HM, Jagger C, et al. Grip strength across the life course: normative data from twelve British studies. PLoS ONE. 2014;9(12):e113637.
46 McQuiddy VA, Scheerer CR, Lavalley R, McGrath T, Lin L. Normative Values for Grip and Pinch Strength for 6- to 19-Year-Olds. Arch Phys Med Rehabil. 2015;96(9):1627–33.
47. Salthouse TA. Why are there different age relations in cross-sectional and longitudinal comparisons of cognitive functioning? Curr Dir Psychol Sci. 2014;23(4):252–6.
48. Gur RC, Richard J, Hughett P, Calkins ME, Macy L, Bilker WB, Brensinger C, Gur RE. A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: standardization and initial construct validation. J Neurosci Methods. 2010;187(2):254–62.
49. Roalf DR, Gur RC, Almasy L, Richard J, Gallagher RS, Prasad K, Wood J, Pogue-Geile MF, Nimgaonkar VL, Gur RE. Neurocognitive performance stability in a multiplex multigenerational study of schizophrenia. Schizophr Bull. 2013;39(5):1008–17.
50. Moore TM, Reise SP, Gur RE, Hakonarson H, Gur RC. Psychometric properties of the Penn Computerized Neurocognitive Battery. Neuropsychology. 2015;29(2):235–46.
51. Thorgusen SR, Suchy Y, Chelune GJ, Baucom BR. Neuropsychological Practice Effects in the Context of Cognitive Decline: Contributions from Learning and Task Novelty. J Int Neuropsychol Soc. 2016;22(4):453–66.
52. Schulte T, Mueller-Oehring EM, Rosenbloom MJ, Pfefferbaum A, Sullivan EV. Differential effect of HIV infection and alcoholism on conflict processing, attentional allocation, and perceptual load: evidence from a Stroop Match-to-Sample task. Biol Psychiatry. 2005;57(1):67–75.
53. Lannoy S, Pfefferbaum A, Le Berre AP, Thompson WK, Brumback T, Schulte T, Pohl KM, De Bellis MD, Nooner KB, Baker FC et al: Growth trajectories of cognitive and motor control in adolescence: How much is development and how much is practice? Neuropsychology. 2021;36(1):44–54.
54. Brown SA, Myers MG, Lippke L, Tapert SF, Stewart DG, Vik PW. Psychometric evaluation of the Customary Drinking and Drug Use Record (CDDR): a measure of adolescent alcohol and drug involvement. J Stud Alcohol. 1998;59(4):427–38.
55. Pfefferbaum A, Kwon D, Brumback T, Thompson WK, Cummins K, Tapert SF, Brown SA, Colrain IM, Baker FC, Prouty D, et al. Altered Brain Developmental Trajectories in Adolescents After Initiating Drinking. Am J Psychiatry. 2018;175(4):370–80.
56. Nichols BN, Pohl KM. Neuroinformatics Software Applications Supporting Electronic Data Capture, Management, and Sharing for the Neuroimaging Community. Neuropsychol Rev. 2015;25(3):356–68.
57. Rohlfing T, Cummins K, Henthorn T, Chu W, Nichols BN. N-CANDA data integration: anatomy of an asynchronous infrastructure for multi-site, multi-instrument longitudinal data capture. J Am Med Inform Assoc. 2014;21(4):758–62.
58. Sullivan EV, Shear PK, Zipursky RB, Sagar HJ, Pfefferbaum A. A deficit profile of executive, memory, and motor functions in schizophrenia. Biol Psychiat. 1994;36(10):641–53.
59. Wood SN. Thin-plate regression splines. J R Stat Soc (B). 2003;65:95–114.
60. Wood SN. Stable and efficient multiple smoothing parameter estimation for generalized additive models. J Am Stat Assoc. 2004;99:673–86.
61. Salthouse TA. Influence of age on practice effects in longitudinal neurocognitive change. Neuropsychology. 2010;24(5):563–72.
62. Satterthwaite TD, Connolly JJ, Ruparel K, Calkins ME, Jackson C, Elliott MA, Roalf DR, Ryan Hopsona KP, Behr M, Qiu H, et al. The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth. Neuroimage. 2016;124(Pt B):1115–9.
63. Schaie KW, Willis SL, Pennak S. An Historical Framework for Cohort Differences in Intelligence. Res Hum Dev. 2005;2(1–2):43–67.

Publisher's Note