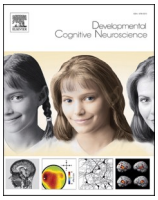
Developmental Cognitive Neuroscience 54 (2022) 101078

Contents lists available at ScienceDirect

Developmental Cognitive Neuroscience

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



Age-related changes and longitudinal stability of individual differences in ABCD Neurocognition measures

Andrey P. Anokhin [a,*], Monica Luciana [b], Marie Banich [c], Deanna Barch [a], James M. Bjork [d], Marybel R. Gonzalez [e], Raul Gonzalez [f], Frank Haist [e], Joanna Jacobus [e], Krista Lisdahl [g], Erin McGlade [h], Bruce McCandliss [i], Bonnie Nagel [j], Sara Jo Nixon [k], Susan Tapert [e], James T. Kennedy [a], Wesley Thompson [e]

[a] *Washington University in St. Louis, USA*
[b] *University of Minnesota, USA*
[c] *University of Colorado Boulder, USA*
[d] *Virginia Commonwealth University, USA*
[e] *University of California San Diego, USA*
[f] *Florida International University, USA*
[g] *University of Wisconsin-Milwaukee, USA*
[h] *The University of Utah, USA*
[i] *Stanford University, USA*
[j] *Oregon Health & Science University, USA*
[k] *University of Florida, USA*

ARTICLE INFO

Keywords:
Neurocognition
Longitudinal
Development
Test-retest reliability

ABSTRACT

Temporal stability of individual differences is an important prerequisite for accurate tracking of prospective relationships between neurocognition and real-world behavioral outcomes such as substance abuse and psychopathology. Here we report age-related changes and longitudinal test-retest stability (TRS) for the Neurocognition battery of the Adolescent Brain and Cognitive Development (ABCD) study, which included the NIH Toolbox (TB) Cognitive Domain and additional memory and visuospatial processing tests administered at baseline (ages 9–11) and two-year follow-up. As expected, performance improved significantly with age, but the effect size varied broadly, with Pattern Comparison and the Crystallized Cognition Composite showing the largest age-related gain (Cohen's d:.99 and.97, respectively). TRS ranged from fair (Flanker test: r = 0.44) to excellent (Crystallized Cognition Composite: r = 0.82). A comparison of longitudinal changes and cross-sectional age-related differences within baseline and follow-up assessments suggested that, for some measures, longitudinal changes may be confounded by practice effects and differences in task stimuli or procedure between baseline and follow-up. In conclusion, a subset of measures showed good stability of individual differences despite significant age-related changes, warranting their use as prospective predictors. However, caution is needed in the interpretation of observed longitudinal changes as indicators of neurocognitive development.

1. Introduction

A key goal of developmental cognitive neuroscience is to evaluate longitudinal changes in neurocognitive functioning. Crucial to understanding the neurodevelopment underlying behavioral disorders, intellectual disability or mental illness is the application of robust assessments of neurocognitive function. However, test-retest reliability of neurocognitive phenotypes puts a critical constraint on the ability to detect meaningful associations with other variables (Hedge et al., 2018; Kanyongo et al., 2007; Vul et al., 2009). A critical issue when assessing stability of a cognitive task performance *across a period of broad cognitive development* arises when potential practice effects add to or interact with global improvements in cognitive function, which themselves might occur at different rates in different participants (Sullivan et al., 2017). Of vital importance is disentangling the effects and relative impacts of practice with aging and related cognitive development. Longitudinal

* Correspondence to: Department of Psychiatry, Washington University School of Medicine, 660S. Euclid Ave, Campus Box 8134, St. Louis, MO 63105, USA.
E-mail address: andrey@wustl.edu (A.P. Anokhin).

https://doi.org/10.1016/j.dcn.2022.101078
Received 15 July 2021; Received in revised form 23 December 2021; Accepted 26 January 2022
Available online 28 January 2022

studies of neurocognitive development that utilize well powered normative samples to enable such comparisons are scarce.

1.1. ABCD study and Neurocognition battery overview

Longitudinal data from the Adolescent Brain and Cognitive Development Study (ABCD Study®) are now available to address and evaluate these methodological issues. The ABCD study is a national consortium that includes 21 data collection sites throughout the United States engaged in a 10 year prospective study of over 11,000 children first assessed at ages 9–10. ABCD assessments are comprehensive and include neuroimaging (structural and functional MRI), a battery of neurocognitive tasks, self- and parental reports about a broad range of behaviors and environmental exposures, and collection of biospecimens. The long-term goals of ABCD include a detailed characterization of adolescent neurocognitive development, identification of risk factors and prospective predictors for future health-related outcomes such as substance use and neuropsychiatric disorders, and elucidation of the effect of various environmental exposures such as substance use as on neurocognitive development (Luciana et al., 2018).

ABCD longitudinal data can address the important question of long-term, longitudinal stability of individual difference in neurocognition, in light of its cohort selection and longitudinal pacing. Notably, the wide age span of entry into the study (the youngest 9-year-olds thru the oldest 10-year olds) relative to the ~2 year span between assessments affords a means to disentangle age effects from practice effects on performance. Temporal stability of individual differences is an important prerequisite for developmental research focused on prospective relationships between neurocognition and real-world behavioral outcomes such as substance abuse and psychopathology because such research relies (often implicitly) on the assumption that neurocognitive "markers", "endophenotypes", and "predictors" represent stable traits (Enkavi et al., 2019; Miller and Rockstroh, 2013).

1.2. Previous studies of Test-Retest Stability (TRS) of neurocognition measures

Previous studies of behavioral and cognitive tasks suggest that robust and reproducible within-subject experimental effects (such as Flanker or Stroop effects) do not necessarily guarantee reliability of individual differences. Hedge et al. (2018) evaluated test-retest reliability (TRR) of seven commonly used neurocognitive tasks with a three-week retest interval and found that only a few indices exceeded the conventional threshold (ICC $\geq$ 0.6) for "good/substantial" reliability (Hedge et al., 2018). Another recent study (Enkavi et al., 2019) assessed TRR of performance in a large set of self-regulation measures in 150 adult participants with an average test-retest interval of 111 days. The analyses yielded median reliability of only 0.31, leading the authors to question the ability of behavioral task measures to serve as trait-like measures of individual differences, however, see (Friedman and Banich, 2019) for a somewhat different viewpoint. For the NIH Toolbox cognition battery, short-term (1–3 weeks) TRR in children (ages 3–15; n: 49–66) was very high (ICC: 0.84–0.99) (Weintraub et al., 2013), however, the two-year longitudinal stability in another sample of children (ages 9–15; n = 118) was substantially lower (ICC: 0.31–0.76) (Taylor et al., 2020). Performance on tests from the Neuropsychological Battery included in the NIH Study of Normal Brain Development showed a wide range of two-year test-retest stability estimates, with IQ and its component measures showing the highest stability (r = 0.81) (Waber et al., 2012). Importantly, previous studies consistently reported low reliability of task scores based on reaction time (RT) difference measures, e.g. RT costs in interference tasks, in contrast to good reliability of the mean RT in individual task conditions, which results in a trade-off between "process purity" of difference measures and reliability of mean RT measures (Draheim et al., 2021; Paap and Sawi, 2016). Recently, Draheim et al. (2021) proposed to address this problem by developing novel accuracy-based performance measures and demonstrated that such measures have higher reliability and validity compared with RT-based measures and are therefore more suitable for individual differences research. Overall, previous research suggests that TRS of neurocognitive performance measures can vary broadly across tasks and samples.

1.3. Practice effects

One potential shortcoming of repeated assessments in longitudinal studies is the possibility of misinterpreting age-related changes as "true" developmental changes, when these differences may be driven by practice effects. For example, analyses of data from the National Consortium on Alcohol and NeuroDevelopment in Adolescence (NCANDA) project examined factors affecting change in scores on 16 neuropsychological test composites over one year in 568 adolescents and suggested that performance gain was mainly attributable to testing experience (practice) with little contribution from predicted developmental effects (Lannoy et al., 2021; Sullivan et al., 2017). Another study of adult participants tested twice with an average retest interval of 2.5 years also showed that practice effects can positively bias the longitudinal trends, but the size of practice effects depended on the age of participants (Salthouse, 2010). These studies underscore the importance of accounting for possible practice effects in the interpretation of age-related changes in longitudinal studies of neurocognition. However, In the NIH Study of Normal Brain Development involving longitudinal assessments of children aged 6–18 years at baseline and a two-year retest interval, only few of the tests from the Neuropsychological Battery showed practice effects, and effect size estimates were small (Waber et al., 2012).

The relative quantification of TRS and practice effects assumes that the same assessment forms or variants are utilized from one assessment wave to the next. Within ABCD, however, retesting of several abilities, particularly explicit memory functions, necessitated the use of alternate forms over time to avoid carryover effects from the prior administration (see Methods). Thus, the interpretation of retest stability from these measures is more complex (Sullivan et al., 2017).

1.4. Aims of the study

The present report focuses on ABCD neurocognitive tasks data collected during the baseline assessment (ages 9–10) and at two-year follow-up (ages 11–12). Our analyses pursued two major aims: first, we evaluated longitudinal changes in neurocognitive performance over the two-year interval between the baseline and follow-up assessment; second, we assessed longitudinal stability of individual differences in task performance. We addressed the following questions/hypotheses:

1) Are there longitudinal changes in neurocognition over a two-year interval? We expected significant improvements in task performance as indicated by gains in accuracy and decreased reaction times as revealed by both longitudinal comparisons and cross-sectional analyses within each assessment wave.
2) Does the rate of longitudinal changes depend on the age at baseline and/or sex? Based on evidence from previous cross-sectional studies suggesting faster developmental changes in younger children and their subsequent decelerations with age (Korkman et al., 2001; Waber et al., 2007), we expected that younger children would show larger age-related gains in performance.
3) Is there evidence for practice effects that might confound developmental changes assessed longitudinally? Consistent with previous literature (Sullivan et al., 2017) we expected practice effects but anticipated they would vary across tests.
4) Is there evidence of ceiling effects, as performance improves with age? We hypothesized that ceiling effects would be most evident for those tests involving a limited number of trials and/or responses and

that focus on accuracy metrics (versus reaction times) as outcome measures.

5) What is the long-term, longitudinal test-retest stability of individual differences in test performance (i.e. in terms of the rank order of performance across participants)? We expected a broad range of test-retest stability estimates, with the highest stability for a composite measure of cognition, in line with a previous developmental study (Taylor et al., 2020) showing stronger reliability for composite scores compared with individual subtests.

2. Material and methods

2.1. ABCD participants

The present analyses utilized ABCD data from the National Institute of Mental Health National Data Archive (NDA) release 4.0 (https://dx.doi.org/10.15154/1523041) that included the baseline in-person assessments (n = 11,876, mean age ± SD: 9.92 ± 0.62 years, 47.8% female) and the 24-month in-person follow-up assessments completed by the time of data release (n = 10,414, mean age ± SD: 12.00 ± 0.66 years, 47.6% female). Parental consent and assent was obtained in minors participating in the study. The interval between the baseline and follow-up assessments (Mean, SD) was 2.09 ± 0.22 years. All data were subjected to quality control (QC) checks by the ABCD Data Analysis and Informatics Core (DAIRC). Because some cases failed to pass the QC check, data were missing for some participants for individual tests and sample size varies slightly across individual analyses (by less than 1% for most measures).

2.2. Neurocognitive assessments

For a detailed description of the ABCD Neurocognition battery and comprehensive analyses of baseline data, see Luciana et al. (2018) and Thompson et al. (Thompson et al., 2019), respectively. The present report represents a longitudinal extension of this previous work enabled with release of the two-year follow-up data. The present analyses utilized only those measures for which longitudinal assessments were available, i.e. at baseline and two-year follow-up including 5 tests from the NIH Toolbox (Picture Vocabulary, Flanker Inhibitory Control & Attention test, Picture Sequence Memory, Pattern Comparison Processing Speed, Oral Reading Recognition as well as a composite measure of Crystallized Cognition (Akshoomoff et al., 2013; Bleck et al., 2013; Weintraub et al., 2013)). Other tests included the Rey Auditory Verbal Learning Test (RAVLT, Rey, 1964; Strauss et al., 2006; Taylor, 1959), a test of verbal learning and memory including immediate and delayed recall and the Little Man Task (LMT, Acker and Acker, 1982), and a mental rotation test of visuospatial processing (Luciana et al., 2018). For the RAVLT, two alternate forms of test containing different word lists (Forms 1 and 5 as described in Hawkins et al., 2004) were used in the baseline and follow-up assessments, respectively, to mitigate potential practice effects. Electronic versions of the tests were administered on iPad in a supervised laboratory setting (see Luciana et al., 2018 for details). There was also a minor change in the LMT administration procedure (moving the "home button" from the tabletop to the touchscreen) at the beginning of the follow-up wave. For the above tests, we used uncorrected scores because the use of age-corrected and fully corrected scores would preclude meaningful analyses of longitudinal changes and cross-sectional age-related differences.

2.3. Statistical analysis

To examine longitudinal *changes* in performance, we compared task performance at baseline and follow-up assessments using paired t-test. To examine possible effects of age at baseline and sex on the rate of age-related change, we used a mixed-design general linear model (GLM) with a longitudinal repeated-measures factor, Age_L, with 2 levels (baseline and follow-up), and tested for the interaction between Age_L and age at the baseline assessment, Age_B (operationalized as three-month age bins, see below). To examine whether male and female participants differ with respect to the rate of neurocognitive development, we tested for Age_L by sex interaction.

To assess age-related *differences* in task performance within each assessment wave, we used a regression analysis with age at assessment as the independent variable and task performance scores as dependent variables. To facilitate the visualization of cross-sectional age-related trends and comparison between the assessment waves, we grouped participants' age into 18 three-month bins (bin 1: greater or equal to 8.75 to less than 9 years, bin 18: greater or equal than 13–13.25 years).

To assess practice effects, we took advantage of the age overlap between the oldest participants at baseline assessment and the youngest participants at follow-up: age bins 9 and 10 spanning the age interval between 10.75 and 11.25 years included both baseline and follow-up assessments (Fig. 1). To enable direct tests of practice effects, we formed age-matched groups of oldest baseline participants and the youngest follow-up participants (n = 787 and 732, respectively, mean age ± SD for the baseline and follow-up groups: 10.93 ± 0.03 and 10.93 ± 0.08, respectively, t = 0.353, df = 935.4, p = .724). These groups consisted of different individuals, i.e. the group comparison was cross-sectional. Importantly, while these groups were matched by age, they differed with respect to their experience with the tests. The oldest participants at the baseline assessment performed the tests for the first time, whereas the youngest participants at the first follow-up assessments performed the tests for the second time, i.e. were already familiar with the tests and the overall testing situation. As explained in more detail below, this fact was considered in the context of our analyses to attempt to disentangle practice from age. Under the practice effect hypothesis, we expected that subjects from the follow-up assessment who were already familiar with the tests would show superior performance compared with their test-naïve age-matched counterparts from the baseline assessment.

Preliminary analyses showed that these groups differed slightly but significantly with respect to parental education level (a proxy for socioeconomic status, lower in the follow-up group) and hormonal measures (higher testosterone and lower DHEA in the follow-up group). These unexpected differences could potentially confound differences in neurocognitive performance because socioeconomic status is known to be a strong determinant of children's neurocognitive development (Bradley and Corwyn, 2002; Gonzalez et al., 2020; Ursache and Noble, 2016), while hormonal status may be related to developmental differences that may not be fully accounted by chronological age due to individual variability in developmental rate, which can affect cognitive performance (Campbell, 2020; Peper and Dahl, 2013). To rule out potential confounding effects, we included parental education and hormonal status as covariates in the analyses of practice effects, although missing data, primarily in hormone measures, decrease sample sizes for some analyses.

To examine the effect of age-related changes on the range of variance in test scores and possible ceiling effects (compression of score variance on the upper end of the distribution as test performance improves with age) we used descriptive statistics and visualization such as distribution histograms and scatterplots.

Longitudinal test-retest reliability of individual differences was assessed using two measures: Pearson correlations and intraclass correlation coefficient (ICC). Pearson (product-moment) correlation evaluates the consistency of relative ranking of individuals within the group across time and is robust to systematic age-related changes in absolute scores and variance. Although most test-retest studies have been using ICC, Rousson et al. (2002) argue that product-moment correlation is more appropriate for test-retest analysis than ICC. ICC assumes random or arbitrary ordering of the measurements within individual (i.e. measurements are interchangeable), which is true in the case of e.g. different raters, but certainly not so in the test-retest situation where the number

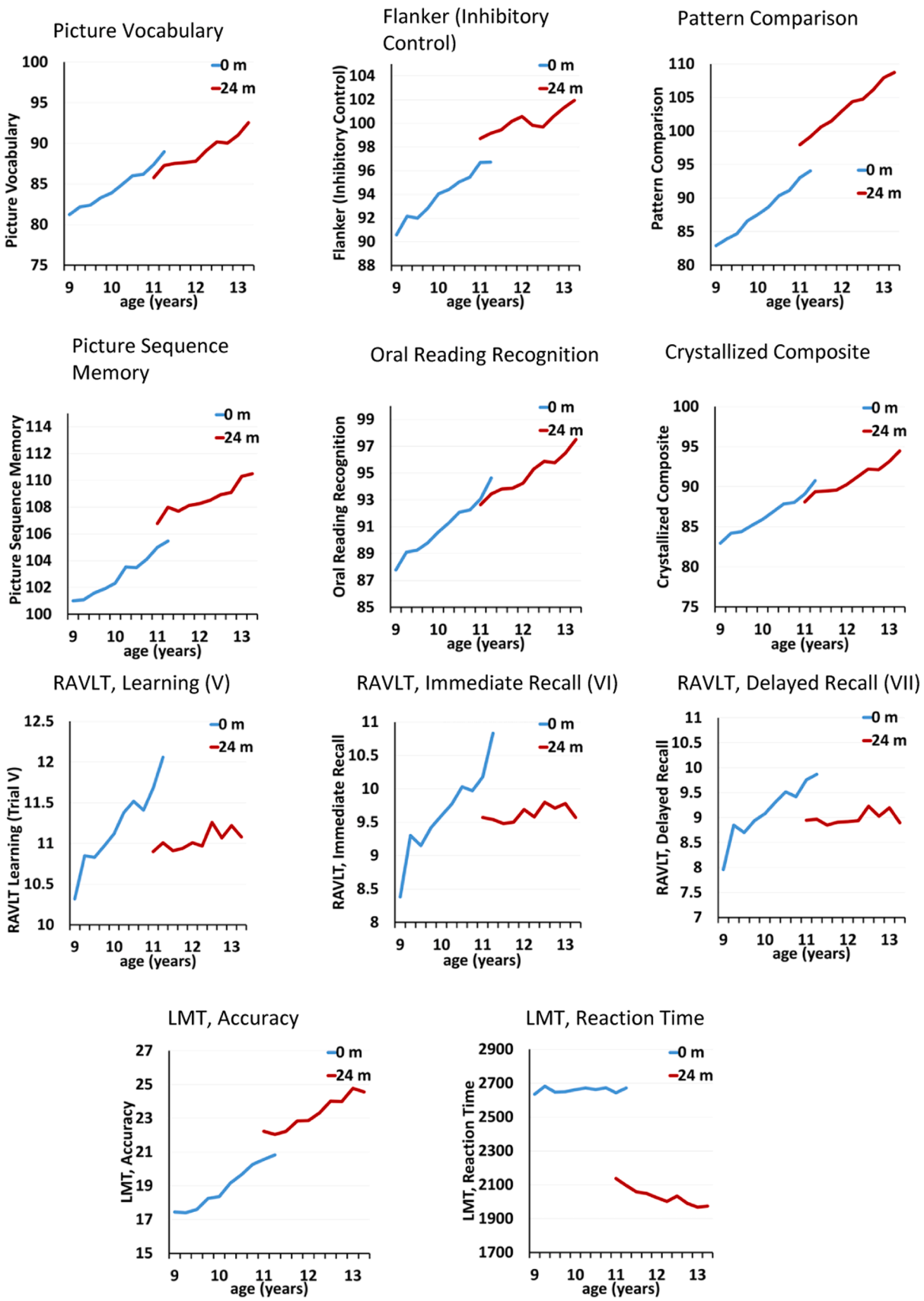


Fig. 1. Longitudinal changes and cross-sectional age-related differences in ABCD Neurocognition measures. Test scores are plotted as a function of age. Horizontal axis: Age (3-month age bins); vertical axis: test score. Blue lines represent baseline data (0 months), and red lines represent two-year follow-up data (24 months). Note an age overlap between the oldest participants at baseline and the youngest participants at the follow-up (bins 9 and 10).

and order of assessments is fixed. This fixed order may result in a systematic "error", i.e. a difference in mean values between the tests due to developmental changes or practice effects. The former is more likely to occur at long retest intervals, whereas the latter is more likely to happen at shorter intervals. The product-moment correlation is not penalized by this "systematic error" and reflects the consistency of individual differences relative to the group mean (i.e. relative ranking), rather than agreement of absolute scores. It is important to note that the product moment correlation is not only robust to systematic shifts in the mean value across measurement occasions, but also to changes in the variance (Rousson et al., 2002). However, since many previous test-retest reliability studies traditionally used ICC (Koo and Li, 2016), we also computed the "consistency" ICC(3,1), according to Shrout and Fleiss (1979) using a 2-way mixed-effects model for a single measurement using SPSS statistical package version 28 (SPSS Inc, Chicago, IL).

3. Results

3.1. Age-related changes

Preliminary analyses of age distribution showed that the age range is relatively broad (9–11 years at baseline and 10.5–14 years at the two-year follow-up), and both distributions are fairly uniform, permitting the analysis of age-related differences within each wave (Supplementary Fig. S1).

Longitudinal analyses showed significant improvement in performance with age across most tests (Table 1; Fig. 1). The only exception was the RAVLT scores for which significant age-related changes were negative.

Cross-sectional analyses within each of baseline and follow-up assessment waves showed significant positive correlations between test performance and age for all assessments, except for the LMT reaction time at the baseline assessment, thus corroborating the results of the longitudinal analysis (Table 2). Significant correlations with age ranged from.11 (RAVLT, Delayed Recall) to.26 (Crystallized Cognition Composite) at baseline and from.02 (RAVLT, Delayed Recall) to.22 (Crystallized Cognition Composite) at follow-up.

Next, we examined factors potentially affecting age related changes, including age at baseline and sex. A general linear models (GLM) analysis showed a significant interaction between the repeated measures (longitudinal) effect of the study wave (baseline versus follow-up) and age at the baseline assessment for most performance variables, except for Pattern Comparison Processing Speed and Picture Sequence Memory from the NIH TB battery (Table 3). This interaction indicates that the longitudinal changes in performance are moderated by baseline age. An illustration of this analysis by example of the Flanker test performance is shown in Fig. S2. Follow-up correlational analysis showed small but consistently negative correlations between the amount of longitudinal change and age at baseline (range of significant correlations: −0.02 to −0.11), indicating that younger participants tended to show larger age-related gains in performance, as expected.

Interactions between the longitudinal changes and sex were mostly non-significant (Table 3), with the exception of all measures of RAVLT, which revealed larger gains in boys, effect sizes were very small (Partial η^2:.001–0.002).

To examine whether the structure of relationships among measures changes with age, we computed Pearson correlations among test scores separately for baseline and follow-up. The size and pattern of these correlations was remarkably similar (Fig. S3). We did not find any systematic increase or decrease in the size of intercorrelations among measures. To evaluate the similarity in the pattern of correlations among measures at baseline and follow-up, we computed a correlation between Fisher-transformed correlations at the two assessment waves, which was r = 0.99, indicating a very high stability of the overall structure of relationships between test scores over the two-year developmental interval.

Table 2
Cross-sectional age-related differences: correlations with age within the baseline and 2-year follow-up assessments.

Assessment	Baseline			Follow-up		
	r	n	p	r	n	p
Picture Vocabulary	0.234	11,728	< 0.001	0.195	9851	< 0.001
Flanker	0.179	11,722	< 0.001	0.091	7934	< 0.001
Pattern Comparison Processing Speed	0.220	11,704	< 0.001	0.211	7896	< 0.001
Picture Sequence Memory	0.113	11,716	< 0.001	0.069	9882	< 0.001
Oral Reading Recognition	0.216	11,714	< 0.001	0.190	9812	< 0.001
Crystallized Cognition Composite	0.257	11,696	< 0.001	0.219	7465	< 0.001
RAVLT, Learning (Trial V)	0.121	11,687	< 0.001	0.035	9921	< 0.001
RAVLT, Immediate Recall(Trial VI)	0.121	11,665	< 0.001	0.029	9872	0.004
RAVLT, Delayed Recall(Trial VII)	0.110	11,611	< 0.001	0.023	9804	0.022
LMT, n correct	0.214	11,538	< 0.001	0.151	9933	< 0.001
LMT, RT correct	-0.009	11,532	0.175	-0.097	9928	< 0.001

Notes: Longitudinal change was computed by subtracting baseline values from 2-year follow-up values, i.e. positive t-values reflect a score increase and vice versa.

Table 1
Longitudinal changes and test-retest stability of individual differences in task performance (ABCD Neurocognition battery, tests administered at baseline and 2-year follow-up). Longitudinal change was computed by subtracting baseline values from follow-up values, i.e. positive t-values reflects increase in test performance and vice versa. Cohen's d indicates the effect size of age-related change. Test-retest stability measures: r: Pearson correlation coefficient; ICC(3,1): intraclass correlation coefficient, consistency type.

Test	Mean±SD		Paired t	df	Cohen's d	p	r	ICC
	Baseline	Follow-up						
Picture Vocabulary	84.8 ± 8.0	89.0 ± 8.5	67.9	9735	0.69	< 0.001	0.73	0.73
Flanker	94.4 ± 8.8	100.1 ± 7.6	57.5	7848	0.65	< 0.001	0.44	0.43
Pattern Comparison Processing Speed	88.4 ± 14.4	103.5 ± 15.1	88.7	7803	1.00	< 0.001	0.48	0.48
Picture Sequence Memory	103.1 ± 12.0	108.7 ± 12.6	41.8	9759	0.42	< 0.001	0.44	0.44
Oral Reading Recognition	91.1 ± 6.8	95.0 ± 6.7	79.6	9685	0.81	< 0.001	0.76	0.76
Crystallized Cognition Composite	86.8 ± 6.9	90.9 ± 7.1	84.3	7369	0.98	< 0.001	0.82	0.82
RAVLT, items learned (trial V)	11.3 ± 2.6	11.1 ± 2.5	-8.3	9772	-0.08	< 0.001	0.43	0.43
RAVLT, Immediate Recall (trial VI)	9.8 ± 3.0	9.6 ± 2.8	-3.6	9707	-0.04	< 0.001	0.47	0.47
RAVLT, Delayed Recall (trial VII)	9.3 ± 3.2	9.0 ± 3.0	-8.9	9595	-0.09	< 0.001	0.51	0.50
LMT, n correct	19.0 ± 5.5	23.3 ± 6.0	75.2	9639	0.77	< 0.001	0.52	0.51
LMT, RT correct	2672.5 ± 464.5	2024.3 ± 480.4	-105.3	9628	-1.07	< 0.001	0.18	0.18

Table 3
Effects of sex and age at the baseline assessment on longitudinal changes in task performance.

Test	Effect	F	df	p	Effect Size (Partial η^2)
Picture Vocabulary	Age_L	1876.319	1,9715	< 0.001	0.162
	Age_L X Age_B	4.891	9,9715	< 0.001	0.005
	Age_L X Sex	0.215	1,9715	0.643	0.000
Flanker	Age_L	1396.705	1,7828	< 0.001	0.151
	Age_L X Age_B	13.525	9,7828	< 0.001	0.015
	Age_L X Sex	3.504	1,7828	0.061	0.000
Pattern Comparison Processing Speed	Age_L	3108.127	1,7783	< 0.001	0.285
	Age_L X Age_B	1.532	9,7783	0.130	0.002
	Age_L X Sex	1.637	1,7783	0.201	0.000
Picture Sequence Memory	Age_L	691.209	1,9739	< 0.001	0.066
	Age_L X Age_B	1.533	9,9739	0.130	0.001
	Age_L X Sex	0.060	1,9739	0.807	0.000
Oral Reading Recognition	Age_L	2769.128	1,9664	< 0.001	0.223
	Age_L X Age_B	7.024	9,9664	< 0.001	0.006
	Age_L X Sex	0.048	1,9664	0.826	0.000
Crystallized Cognition Composite	Age_L	2968.169	1,7349	< 0.001	0.288
	Age_L X Age_B	6.290	9,7349	< 0.001	0.008
	Age_L X Sex	0.125	1,7349	0.724	0.000
RAVLT, Learning (Trial V)	Age_L	30.431	1,9753	< 0.001	0.003
	Age_L X Age_B	5.460	9,9753	< 0.001	0.005
	Age_L X Sex	5.499	1,9753	0.019	0.001
RAVLT, Immediate Recall (Trial VI)	Age_L	3.356	1,9688	0.067	0.000
	Age_L X Age_B	9.692	9,9688	< 0.001	0.009
	Age_L X Sex	16.949	1,9688	< 0.001	0.002
RAVLT, Delayed Recall (Trial VII)	Age_L	27.121	1,9576	< 0.001	0.003
	Age_L X Age_B	6.592	9,9576	< 0.001	0.006
	Age_L X Sex	8.900	1,9576	0.003	0.001
LMT, n correct	Age_L	2355.369	1,9620	< 0.001	0.197
	Age_L X Age_B	1.924	9,9620	0.044	0.002
	Age_L X Sex	0.299	1,9620	0.584	0.000
LMT, RT correct	Age_L	4501.577	1,9609	< 0.001	0.319
	Age_L X Age_B	3.342	9,9609	< 0.001	0.003
	Age_L X Sex	0.576	1,9609	0.448	0.000

Notes: Age_L is a longitudinal, within-subject factor with two levels (baseline, follow-up); Age_B is age at baseline (a between-subject factor with 10 levels corresponding to 3-month age bins). A significant Age_L X Age_B interaction indicates that the rate of longitudinal change varies as a function of age at baseline. A significant Age_L X Sex interaction indicates that the rate of longitudinal change differs between girls and boys.

3.2. Practice effects

The pattern of longitudinal changes and cross-sectional differences (Fig. 1) suggested possible practice effects for five performance variables, including three NIH TB tests (Flanker, Pattern Comparison Processing Speed, and Picture Sequence Memory) and the Little Man Task (LMT) accuracy and reaction time. Specifically, the youngest participants at the follow-up assessment showed markedly better performance than their age-matched counterparts at their baseline assessment (Table 4). In the absence of practice effects one would expect a perfect alignment of the end of the baseline age dependency curve and the beginning of the follow-up age dependency curve (blue and red curves in Fig. 1, respectively). However, quite unexpectedly, there were also significant differences in the opposite direction for all three RAVLT variables, indicating that the youngest follow-up participants who had already had experience with the test performed *worse* than their age-matched counterparts at baseline who performed the test for the first time. This puzzling "negative practice effect" may be due to a change in the experimental procedure from baseline to follow-up assessments (see Discussion for more details).

3.3. Ceiling effects

Picture Sequence Memory and the Little Man Task accuracy score showed significant ceiling effects (score compression in the upper end of the distribution), which became more prominent in the follow-up data (Fig. 2).

3.4. Relationships between developmental changes across neurocognitive domains

To examine whether developmental changes in different neurocognitive processes are correlated, i.e. individuals showing steeper changes in one domain also show steeper changes in others and vice versa, we computed correlations among change scores (the difference between baseline and follow-up scores) for all tests. A pattern of similar rates of change across tests would suggest a general factor of cognitive development, whereas variability in these patterns would be consistent

Table 4
Cross-sectional comparison of age-matched subjects from baseline and follow-up assessments.

Test	Mean (n)		Difference	p	η^2_P
	Baseline	Follow-up			
Picture Vocabulary	88.203 (736)	87.364 (334)	-0.839	.105	.002
Flanker	97.200 (736)	99.140 (341)	1.940	< 0.001	.012
Pattern Comparison	93.163 (734)	99.037 (340)	5.874	< 0.001	.033
Picture Sequence Memory	105.258 (737)	108.185 (340)	2.926	< 0.001	.011
Oral Reading Recognition	93.719 (735)	93.291 (333)	-0.428	.337	.001
Crystallized Cognition Composite	89.829 (734)	89.265 (340)	-0.564	.195	.002
RAVLT, Learning (Trial V)	11.823 (738)	11.071 (338)	-0.752	< 0.001	.020
RAVLT, Immediate Recall (Trial VI)	10.415 (736)	9.578 (338)	-0.837	< 0.001	.018
RAVLT, Delayed Recall (Trial VII)	9.802 (734)	8.915 (334)	-0.887	< 0.001	.018
LMT, n correct	20.900 (725)	22.872 (342)	1.971	< 0.001	.025
LMT, RT correct	2651.748 (725)	2141.995 (342)	-509.752	< 0.001	.216

Notes: Difference between the groups was computed by subtracting baseline values from the follow-up values, i.e. positive values reflects larger scores in the youngest subjects in the follow-up assessment compared to the age-matched oldest subjects in the baseline assessment and vice versa. Average parental education, DHEA, and testosterone levels were included as SES and developmental covariates, respectively. The LMT Reaction time (RT) is inverse measure of performance with smaller values indicating higher speed. η^2_P (partial eta squared) indicates the effects size.

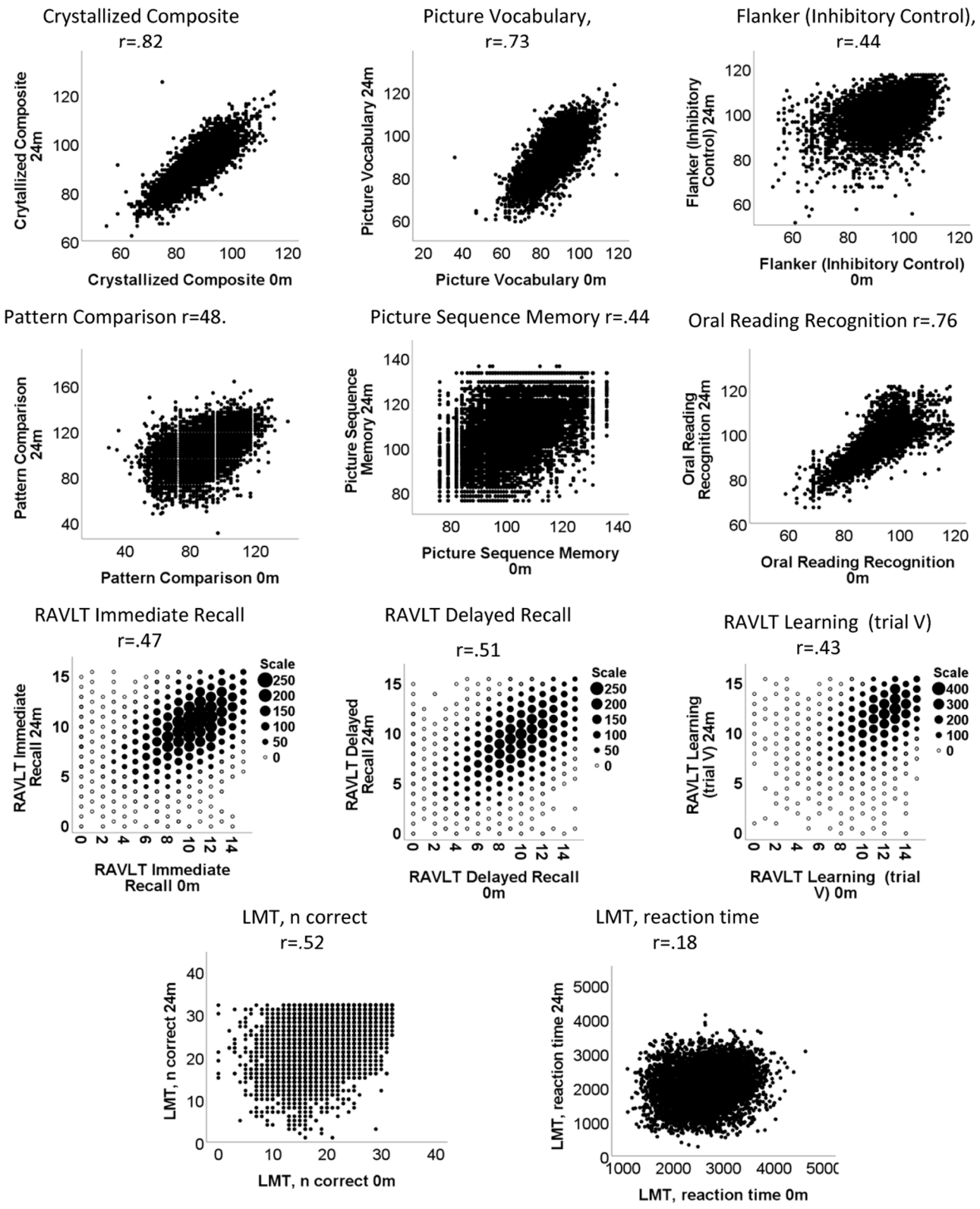


Fig. 2. Scatterplots of longitudinal test-retest correlations. Horizontal and vertical axes represent test scores at baseline and follow-up, respectively.

with test-specific improvement. With the exception of predictably high correlations between changes in the NIH TB Crystallized Cognition Composite measure and its constituents (Picture Vocabulary, r = 0.79 and Oral Reading, r = 0.63), as well as the three RAVLT measures (0.50–0.54), correlations among changes in different test performance measures were low, with the largest correlation of r = 0.23 observed between the NIH TB Flanker and Pattern Comparison Speed tasks. All other correlations were less than 0.1.

3.5. Longitudinal test-retest stability

TRR was significant for all tests but ranged from fair (Flanker: r = 0.44) to excellent (Crystallized Cognition Composite: r = 0.82) (Table 1, Fig. 2). The two metrics of longitudinal test-retest stability, Pearson's r and ICC(3,1), showed highly convergent results, with a negligible mean difference of.001 across 11 variables. Average stability across 10 neurocognition variables (excluding the problematic measure LMT RT, see Discussion) was.56, i.e. in the fair range but close to the conventional threshold of.6 for good reliability (Cicchetti, 2016).

4. Discussion

The primary aims of the present analyses were to evaluate longitudinal changes in neurocognition performance over a two-year period in pre-adolescence, here between the baseline and follow-up assessments of the ABCD study sample and also to assess the longitudinal stability of individual differences in task performance.

As expected, test performance showed significant improvement with age, with the exception of RAVLT delayed recall. However, the effect size varied broadly, from Pattern Recognition and Crystallized Composite scores showing the largest age-related gains (d = 1.00 and.98, respectively) to Immediate Recall on the RAVLT (d = -0.04, a very small effect). For most measures, except RAVLT and LMT-RT, cross-sectional age-related differences were highly consistent with longitudinal changes. Overall, there is strong evidence for substantial improvement in neurocognitive performance over the initial two-year period of the study.

The extent of age-related change depended on baseline age for some variables, with younger participants showing greater changes over two years than their older counterparts (Table 3), consistent with an asymptotic relationship of performance to age, with stronger relationships in younger children found in previous studies (Waber et al., 2007, 2012). However, this pattern is not consistent with cross-sectional age-related differences, which showed a largely linear dependence of test performance on age. With assessment currently available at only two time points, it is difficult to make a definitive conclusion about the shape of developmental trajectories. Subsequent longitudinal waves of the ABCD study will allow us to clarify this issue. The effect of sex on age-related changes was significant for some variables, however, the effect sizes were very small and accounted for less than 1% of the total variance, suggesting that for the measures studied here, neurocognition develops largely at the same rate in boys and girls during the age range examined.

Consistent with our expectations and previous studies (Slade et al., 2008; Sullivan et al., 2017), we found evidence for practice effects for several variables including the NIH Toolbox Flanker, Pattern Comparison Processing Speed, and Picture Sequence Memory tests, and both accuracy and reaction time of the Little Man Task (Fig. 1, Table 4). Specifically, participants with prior experience with the tests showed significantly better performance compared with their age-matched counterparts who performed the tests for the first time. The size of the practice effect was comparable with the amount of age-related gain in performance over one year. Consequently, in a longitudinal comparison with a two-year interval between assessments, "developmental" changes can be overestimated by 50%, if the practice effect is not accounted for. A previous study (Sullivan et al., 2017) found even larger practice effects that accounted for most of the observed longitudinal changes in neurocognitive performance over one year. Therefore, for correct interpretation of longitudinal findings, it is essential to distinguish developmental gains in neurocognitive performance from performance improvements due to prior experience with the test. In the ABCD sample, this is made possible by the reasonably broad age range of the longitudinal cohort (approximately 2 years, which is comparable with the 2-year intervals between longitudinal assessments) allowing for the assessment and comparison of both within-subject age-related *changes* and between-subject age-related *differences*. A partial age overlap between baseline and follow-up assessments permitted a direct comparison of subgroups that are of the same age but differ with respect to their prior experience with the test materials and procedure. Notably, two of the NIH Toolbox tests, Picture Vocabulary and Oral Reading Recognition, as well as their derivative, Crystallized Cognition Composite, did not show significant practice effects. This finding is consistent with a previous report of very small to lacking practice effects for a vocabulary test, in contrast to other tests such as speed and memory, in young adults (Salthouse, 2010). Tests based on verbal knowledge may be less affected by repeated testing than executive function tests where certain task performance strategies or skills can be developed during the first test administration.

The LMT average reaction time showed a particularly large practice effect. Although some shortening of reaction time due to prior experience with the test was expected, the effect was unusually large. It is important to note that this effect may be confounded by a change in the test administration procedure, which coincided with the beginning of the follow-up assessment. Administration of the task was shifted from an in-house programming platform to the commercial Inquisit by Millisecond platform. This shift could also explain the unusually low longitudinal test-retest correlation (r = 0.18) for the average reaction time in this task, given that average reaction time measures (unlike RT difference measures) tend to show good stability (Brown et al., 2014; Hedge et al., 2018). Researchers using LMT reaction time are urged to interpret longitudinal results involving this measure with caution.

Importantly, all three measures of RAVLT performance showed a "reverse practice effect", i.e., worse test performance in the youngest participants at the follow-up assessment (who had prior test experience) relative to their age-matched but test-naïve counterparts at baseline. This counterintuitive finding contradicting the practice effect hypothesis may be related to the use of a different (alternate) version of the test at follow-up relative to baseline that was not well matched with respect to difficulty of the test at baseline, resulting in worse performance at follow-up. Multiple forms of the RAVLT have been created and examined in selected studies (Hawkins et al., 2004). The two forms that we utilized (Forms 1 and 5 as described in Hawkins (2004)) have been contrasted in a limited way in adult samples (Crawford et al., 1989; Majdan et al., 1996). Though these authors concluded that the forms are largely equivalent, Form 1 that we used at baseline ("Drum" list) may be marginally easier than Form 5 used at follow up ("Doll" list) (Hawkins et al., 2004), a difference that may be more substantial in children. Therefore, any longitudinal findings involving RAVLT measures in ABCD data should be interpreted with caution. Excluding measures potentially affected by changes in the experimental procedure (RAVLT and LMT reaction time), as well as Crystallized Composite (which is derived from Picture Vocabulary and Oral Reading and thus is not an independent test), the majority (four out of six) neurocognition measures showed a significant positive practice effect, consistent with previous literature showing strong practice effects for similar measures (Slade et al., 2008; Sullivan et al., 2017).

A ceiling effect was evident for Picture Sequence Memory Test and LMT accuracy (Fig. 2), with a score compression at the upper end of the scale in the follow-up data, indicating that many children performed at ceiling. The ceiling effect for the PSMT is consistent with a previous report based on the PING study (Akshoomoff et al., 2014). Since this ceiling effect was almost absent at baseline and emerged at follow-up, it

can be explained by both developmental improvements in neurocognitive performance (as supported by significant cross-sectional age-related differences, Fig. 1 and Table 2) and practice effects (Table 4). This ceiling effect may be further exacerbated in subsequent longitudinal waves of ABCD and should be taken into account in the interpretation of findings involving these measures.

The amount of longitudinal change in test performance showed weak correlations among different tests, suggesting that the rates of developmental changes in specific neurocognitive functions (processing speed, memory, attention, etc.) are relatively independent in the age range studied here. Subsequent longitudinal waves of longitudinal assessments in the ABCD study should allow us to determine whether developmental trajectories for specific functions remain relatively independent over longer developmental periods.

Longitudinal test-retest stability of individual differences ranged from fair (Flanker test: $r = 0.44$) to excellent (Crystallized Cognition composite: $r = 0.82$). Overall, these stability estimates for NIH TB are highly consistent with a recent three-year longitudinal study of NIH TB involving youth aged 9–15 (Taylor, 2020, see comparison with the present results in Fig. S4). Using conventional criteria (Cicchetti, 2016), a majority of the neurocognition measures were in the "fair" range of reliability (0.40–0.59), with only a few measures reaching threshold for "good" (0.60) or "excellent" (0.75) reliability (Table 1). One possible cause of limited reliability/stability may be insufficient construct validity, i.e., task performance fails to capture the targeted latent construct. As noted by Hedge (2018), robust experimental effects do not necessarily translate to optimal methods of studying individual differences. Another factor negatively affecting reliability may be inconsistent task engagement due to poor motivation, distractions, anxiety, etc. Another well-known factor is the number of trials included in the task. The NIH-TB tasks have been designed to minimize administration time and subject burden and thus were well suited for the ABCD study that strived to minimize assessment time across all assessment domains, in order to maintain a comprehensive, multi-disciplinary assessment protocol. However, the number of trials inversely affects both the size of experimental effects such as Flanker or Stroop effects as well as TRR of the summary performance score (Hedge et al., 2018; Rouder and Haaf, 2019).

It is important to note that the apparently lackluster TRS values reported herein reflect long-term, longitudinal stability in a developmental sample with continuing changes in neurocognitive performance, rather than the relatively stable platform of short retest intervals in adult participants. Moreover, the presence of reliable and longitudinally stable individual differences, despite significant systematic changes in neurocognitive performance with age, indicates that many measures in the ABCD neurocognition battery can be utilized in research focused on prospective associations between neurocognition and real-life outcomes such as substance abuse and psychiatric disorders. However, test-retest reliability is an important factor limiting effect sizes of correlations with other variables and should be factored into statistical power calculations (Hedge et al., 2018). The present results support the notion that the presence of significant age-related changes in test scores does not preclude longitudinal stability of individual differences (rank-order stability) and, conversely, high test-retest stability does not mean that performance does not improve with age (Table 1). Thus, the Crystallized Intelligence Composite score shows a steep age-related improvement but, at the same time, this measure shows excellent test-retest stability ($r = 0.82$) over two years.

Finally, analysis of longitudinal changes in the pattern of relationships among neurocognitive measures showed that neither the strength, nor the pattern of intercorrelations change with age. Analysis of relationships among age-related changes in different measures (difference scores) showed that changes in performance on different tests are largely independent.

The present analyses have some important limitations. First, they were restricted to those measures from the neurocognition battery for which longitudinal data (baseline and two-year follow-up) were available, omitting variables that were introduced in the second or third year of the study as well as two NIH Toolbox measures that were not retained for the Year 2 longitudinal assessment. Second, with only two time points available, the outcomes reflect a short period of development and thus do not inform inferences regarding the shape of developmental trajectories. Data collected in subsequent assessment waves of ABCD study should allow researchers to address these questions. Finally, our analyses focused on age-related changes (assessed longitudinally) and age-related differences (assessed cross-sectionally), whereas analyses of various biological and sociodemographic factors potentially contributing to individual differences in neurocognition (Gonzalez et al., 2020) is beyond the scope of the present report.

Future analytic efforts should apply methods that may increase reliability of neurocognition data in ABCD and other datasets, including latent variable approaches, and implement approaches to test scoring that better account for intra-individual trial-by-trial variability, such as the use of hierarchical linear models that model trial-by-trial variation as well as variation across individuals (Rouder and Haaf, 2019).

5. Conclusions

Task performance showed significant improvement over a two-year period across most tests included in the ABCD Neurocognition battery. This improvement was largely mirrored by cross-sectional age-related differences within each longitudinal assessment. Longitudinal test-retest stability of test performance ranged from fair to excellent. There was evidence suggesting significant positive practice effects for several tests (Flanker, Pattern Comparison, Picture Sequence Memory, and the Little Man Task), which has to be accounted for in the analyses of developmental changes. There was also evidence for a ceiling effects for performance in the Picture Sequence Memory and Little Man task accuracy, suggesting a possibility of further score compression in subsequent follow-up assessments. Longitudinal changes in performance on RAVLT and LMT tasks may be affected by alterations of the experimental procedure, which should be taken into account in the analyses involving these tests and interpretation of results.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development (ABCD) Study (abcdstudy.org), held in the NIMH Data Archive (NDA). The present analyses utilized ABCD data from the National Institute of Mental Health National Data Archive (NDA) release 4.0 (https://dx.doi.org/10.15154/1523041).

Acknowledgments

Research reported in this publication was supported by the grant U01 DA041120-06 from the National Institute on Drug Abuse (NIDA) and the grant R01HD083614 from Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors acknowledge organizational and technical support by the ABCD staff. The authors also acknowledge the generous giving of time by the study participants and their families.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.dcn.2022.101078.

References

Acker, W., Acker, C., 1982. Bexley Maudsley Automated Psychological Screening and Bexley Maudsley Category Sorting Test Manual. NFER-Nelson Publishing, Windsor, Great Britain.

Akshoomoff, N., Beaumont, J.L., Bauer, P.J., Dikmen, S.S., Gershon, R.C., Mungas, D., Slotkin, J., Tulsky, D., Weintraub, S., Zelazo, P.D., Heaton, R.K., 2013. VIII. NIH Toolbox Cognition Battery (CB): composite scores of crystallized, fluid, and overall cognition. Monogr. Soc. Res. Child Dev. 78, 119–132.

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.

Bleck, T.P., Nowinski, C.J., Gershon, R., Koroshetz, W.J., 2013. What is the NIH toolbox, and what will it mean to neurology? Neurology 80, 874–875.

Bradley, R.H., Corwyn, R.F., 2002. Socioeconomic status and child development. Annu. Rev. Psychol. 53, 371–399.

Brown, H.M., Eley, T.C., Broeren, S., MacLeod, C., Rinck, M., Hadwin, J.A., Lester, K.J., 2014. Psychometric properties of reaction time based experimental paradigms measuring anxiety-related information-processing biases in children. J. Anxiety Disord. 28, 97–107.

Campbell, B., 2020. DHEAS and human development: an evolutionary perspective. Front. Endocrinol. 11.

Cicchetti, D., 2016. Developmental Psychopathology, Theory and Method. John Wiley & Sons, Incorporated, New York, United States.

Crawford, J.R., Stewart, L.E., Moore, J.W., 1989. Demonstration of savings on the AVLT and development of a parallel form. J. Clin. Exp. Neuropsychol. 11, 975–981.

Draheim, C., Tsukahara, J.S., Martin, J.D., Mashburn, C.A., Engle, R.W., 2021. A toolbox approach to improving the measurement of attention control. J. Exp. Psychol. Gen. 150, 242–275.

Enkavi, A.Z., Eisenberg, I.W., Bissett, P.G., Mazza, G.L., MacKinnon, D.P., Marsch, L.A., Poldrack, R.A., 2019. Large-scale analysis of test-retest reliabilities of self-regulation measures. Proc. Natl. Acad. Sci. USA 116, 5472–5477.

Friedman, N.P., Banich, M.T., 2019. Questionnaires and task-based measures assess different aspects of self-regulation: both are needed. Proc. Natl. Acad. Sci. USA 116, 24396–24397.

Gonzalez, M.R., Palmer, C.E., Uban, K.A., Jernigan, T.L., Thompson, W.K., Sowell, E.R., 2020. Positive economic, psychosocial, and physiological ecologies predict brain structure and cognitive performance in 9-10-year-old children. Front. Hum. Neurosci. 14, 578822-578822.

Hawkins, K.A., Dean, D., Pearlson, G.D., 2004. Alternative forms of the Rey Auditory Verbal Learning Test: a review. Behav. Neurol. 15, 99–107.

Hedge, C., Powell, G., Sumner, P., 2018. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Behav. Res Methods 50, 1166–1186.

Kanyongo, G.Y., Brook, G.P., Kyei-Blankson, L., Gocmen, G., 2007. Reliability and statistical power: how measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. J. Mod. Appl. Stat. Methods 6, 81–90.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15, 155–163.

Korkman, M., Kemp, S.L., Kirk, U., 2001. Effects of age on neurocognitive measures of children ages 5 to 12: a cross-sectional study on 800 children from the United States. Dev. Neuropsychol. 20, 331–354.

Lannoy, S., Pfefferbaum, A., Le Berre, A.P., Thompson, W.K., Brumback, T., Schulte, T., Pohl, K.M., De Bellis, M.D., Nooner, K.B., Baker, F.C., Prouty, D., Colrain, I.M., Nagel, B.J., Brown, S.A., Clark, D.B., Tapert, S.F., Sullivan, E.V., Müller-Oehring, E. M., 2021. Growth trajectories of cognitive and motor control in adolescence: how much is development and how much is practice? Neuropsychology.

Luciana, M., Bjork, J.M., Nagel, B.J., Barch, D.M., Gonzalez, R., Nixon, S.J., Banich, M.T., 2018. Adolescent neurocognitive development and impacts of substance use: overview of the adolescent brain cognitive development (ABCD) baseline neurocognition battery. Dev. Cogn. Neurosci. 32, 67–79.

Majdan, A., Sziklas, V., Jones-Gotman, M., 1996. Performance of healthy subjects and patients with resection from the anterior temporal lobe on matched tests of verbal and visuoperceptual learning. J. Clin. Exp. Neuropsychol. 18, 416–430.

Miller, G.A., Rockstroh, B., 2013. Endophenotypes in psychopathology research: where do we stand? Annu Rev. Clin. Psychol. 9, 177–213.

Paap, K.R., Sawi, O., 2016. The role of test-retest reliability in measuring individual and group differences in executive functioning. J. Neurosci. Methods 274, 81–93.

Peper, J.S., Dahl, R.E., 2013. The teenage brain: surging hormones—brain-behavior interactions during puberty. Curr. Dir. Psychol. Sci. 22, 134–139.

Rey, A., 1964. L 'examen clinique en psychologie [Clinical tests in psychology]. Presses Universitaires de France, Paris.

Rouder, J.N., Haaf, J.M., 2019. A psychometrics of individual differences in experimental tasks. Psychon. Bull. Rev. 26, 452–467.

Rousson, V., Gasser, T., Seifert, B., 2002. Assessing intrarater, interrater and test-retest reliability of continuous measurements. Stat. Med. 21, 3431–3446.

Salthouse, T.A., 2010. Influence of age on practice effects in longitudinal neurocognitive change. Neuropsychology 24, 563–572.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428.

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.

Strauss, E., Sherman, E.M.S., Spreen, O., 2006. A Compendium of Neuropsychological Tests, third ed. Oxford University Press, New York, New York.

Sullivan, E.V., Brumback, T., Tapert, S.F., Prouty, D., Fama, R., Thompson, W.K., Brown, S.A., Cummins, K., Colrain, I.M., Baker, F.C., Clark, D.B., Chung, T., De Bellis, M.D., Hooper, S.R., Nagel, B.J., Nichols, B.N., Chu, W., Kwon, D., Pohl, K.M., Pfefferbaum, A., 2017. 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. 24, 72–83.

Taylor, B.K., Frenzel, M.R., Eastman, J.A., Wiesman, A.I., Wang, Y.P., Calhoun, V.D., Stephen, J.M., Wilson, T.W., 2020. Reliability of the NIH toolbox cognitive battery in children and adolescents: a 3-year longitudinal examination. Psychol. Med. 1–10.

Taylor, E.M., 1959. The Appraisal of Children with Cerebral Deficits. Harvard University Press, Cambridge, MA.

Thompson, W.K., Barch, D.M., Bjork, J.M., Gonzalez, R., Nagel, B.J., Nixon, S.J., Luciana, M., 2019. The structure of cognition in 9 and 10 year-old children and associations with problem behaviors: findings from the ABCD study's baseline neurocognitive battery. Dev. Cogn. Neurosci. 36, 100606.

Ursache, A., Noble, K.G., 2016. Neurocognitive development in socioeconomic context: multiple mechanisms and implications for measuring socioeconomic status. Psychophysiology 53, 71–82.

Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect. Psychol. Sci.: J. Assoc. Psychol. Sci. 4, 274–290.

Waber, D.P., Moor, De, Forbes, C., Almli, P.W., Botteron, C.R., Leonard, K.N., Milovan, G., Paus, D., Rumsey, J, T., 2007. The NIH MRI study of normal brain development: performance of a population based sample of healthy children aged 6 to 18 years on a neuropsychological battery. J. Int. Neuropsychol. Soc. 13, 729–746.

Waber, D.P., Forbes, P.W., Almli, C.R., Blood, E.A., 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, 179–190.

Weintraub, S., Dikmen, S.S., Heaton, R.K., Tulsky, D.S., Zelazo, P.D., Bauer, P.J., Carlozzi, N.E., Slotkin, J., Blitz, D., Wallner-Allen, K., Fox, N.A., Beaumont, J.L., Mungas, D., Nowinski, C.J., Richler, J., Deocampo, J.A., Anderson, J.E., Manly, J.J., Borosh, B., Havlik, R., Conway, K., Edwards, E., Freund, L., King, J.W., Moy, C., Witt, E., Gershon, R.C., 2013. Cognition assessment using the NIH Toolbox. Neurology 80, S54–S64.