More alike than different: Novel methods for measuring and modeling executive function development

Jessica Wise Younger (✉ jessica@youngers.org)

University of California, San Francisco

Kristine D. O'Laughlin

University of California, San Francisco

Joaquin A. Anguera

University of California, San Francisco

Silvia A. Bunge

University of California, Berkeley

Emilio E. Ferrer

University of California, Davis

Fumiko Hoeft

University of Connecticut

Bruce D. McCandliss

Stanford University

Jyoti Mishra

University of California, San Diego

Miriam Rosenberg-Lee

Rutgers University - Newark

Adam Gazzaley

University of California, San Francisco

Melina R. Uncapher

University of California, San Francisco

Research Article

Keywords: Executive functions, development, measurement, middle childhood

Posted Date: November 23rd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1032581/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Executive functions (EFs) are linked to positive outcomes across the lifespan. Yet, methodological challenges have prevented rigorous understanding of the precise ways EFs are organized in childhood and how they develop over time. We introduce novel methods to address these challenges for both measuring and modeling EFs using a large, accelerated longitudinal dataset from a diverse sample of students in middle childhood (approximately ages 8 to 14; $N = 1,286$). Adaptive assessments allowed us to equate EF challenge across ages and a data-driven, network analytic approach revealed the evolving diversity of EFs while accounting for their unity. Our results suggest EF organization stabilizes around age 10, but continues refining through at least age 14. This approach brings new precision to EFs' development by removing interpretative ambiguities associated with previous methodologies. By improving EF *measurement*, the field can move towards improving EF *training*, to provide a strong foundation for students' success.

Main

A critical set of skills that consistently predict positive outcomes in school and across the lifespan are executive functions (EFs), which comprise a variety of cognitive abilities that enable agency over one's attention (for review see e.g.,^{1,2}). Thus, understanding how EFs emerge and change across development is critical to understanding how we might support their growth during periods of vulnerability and opportunity (from early childhood through and into adulthood). Like most complex cognitive processes, defining and measuring EFs has been complicated, and has not yielded a consistent taxonomy of EFs (see e.g.,³). EFs are often reductively defined as a set of separable but interrelated components, and much of the developmental literature has examined components as separate constructs. Yet, neural data suggest EF components are more alike than different, relying on similar networks, rather than operating as distinct, independent processes (see⁴ for review). Indeed, the neuroimaging field has developed sophisticated methods for interrogating the complex, dynamical relationships between components of neural systems, yet, methods for measuring and modeling such dynamical interactions using behavioral input have lagged behind. We propose that to close this gap between models of neural and behavioral data, new methods must be employed that enable the measurement and modeling of how cognitive processes, including EFs, function in concert to achieve a specific goal⁵. Here we introduce two novel methods, one for measurement and one for modeling, to understand how EFs manifest over development with data we collected in a large, accelerated longitudinal study with a diverse sample of students over two years. First, we show how a novel, adaptive EF assessment battery solves previous challenges to *measuring* EFs consistently. We then *model* these data using novel network analytic techniques to account for what is common across EFs to reveal a clear timeline of EF development during middle childhood, a particularly understudied period.

A key developmental theory posits that EFs begin as a unitary construct in early childhood, and the differentiation of specialized components over time is initiated by experience, to become the multi-dimensional construct observed by young adulthood^{6,7}. This *differentiation hypothesis* aligns with neural developmental evidence showing increased specialization of the neural systems supporting EFs⁸. However, behavioral tests of this hypothesis to-date have largely taken reductionist approaches, modeling EF components as separate processes and failing to adequately account for the commonalities between

components. This approach has revealed no consistent patterns across decades of studies regarding the number of clearly distinguishable components at each age (see ⁹ for review). The inconsistencies in the extant literature call for a paradigm shift in approach to both the measurement and modeling of EF performance to move beyond reductionist views of EF and toward treating them as a dynamic interconnected network of skills. Next, we outline the critical factors that could comprise such an approach and offer evidence in support of the promise of such an approach.

Measuring EFs

To reveal the true developmental trajectory of EFs, we first need to measure EFs with assessments that are robust across developmental stages and assessment sessions. Much of the prior cross-sectional and longitudinal work has been confounded by (1) the use of the *same* tasks across age ranges, which results in floor or ceiling effects in performance if the challenge level is not adjusted appropriately, or (2) the use of *different* tasks with different age groups (as reviewed in e.g., ¹⁰). Not only do tasks need to be comparable across age and ability, but EF assessments also need to be *repeatable* over multiple timepoints so that developmental progress can be measured within subjects in ways that do not suffer from practice or ceiling effects. Adaptive methods that use tasks that dynamically adjust to an individual's appropriate challenge level on a trial-by-trial basis, presents a compelling and simple solution to this pernicious problem^{11,12}. Indeed, prior work with pediatric populations suggests that highly engaging assessments with adapting challenge algorithms can reveal phenotypic differences between clinical and neurotypical populations, even when group characteristics are highly variable¹³.

We further need *multidimensional assessments* to disentangle what EFs share in common from what they uniquely contribute to performance, to ensure each component is measured validly and reliably. Any single task used to assess a component of EF will necessarily involve processes not related to EFs (e.g., visual processing, motor response), or may be related to multiple EF components, both of which will result in measurement impurity¹⁴. To address this impurity, researchers can collect multiple measures of each hypothesized component of EF, leveraging the commonalities across tasks to extract information about EF skills, and reducing the contribution of idiosyncratic skill related to any individual task. Thus, methods that use multiple indicators to measure each hypothesized component of EF are critical for a robust and reliable understanding of how EFs develop over time.

Here we introduce a novel assessment tool—Adaptive Cognitive Evaluation (ACE)—that addresses these robust measurement needs. ACE is a battery of assessments that taps multiple EFs through several different tasks. Importantly, each task incorporates adaptive algorithms, allowing the repeated measurement of EFs across multiple timepoints, using the same tasks in different age groups without running into floor or ceiling limitations.

Modeling EFs

Determining the true developmental trajectory of EFs requires more than robust and reliable measurements—these measurements must also be modeled using methods that can account for the multidimensional nature of EFs. To date, modeling approaches have largely focused on identifying when EFs differentiate over

development, and have not fleshed out how the commonalities among EFs develop and interact. The use of latent variable analysis in particular has been the most common approach to evaluating the changing organization of EFs over development¹⁵. While this modeling technique is attractive because it presents a possible solution to the measurement impurity problem, it is also reductionist as it assumes EFs are distinct, largely non-overlapping constructs. This assumption stands in contrast with the predictions of the differentiation hypothesis which proposes that, while EF components become increasingly separable over time, components do not become *completely* distinct. The persistence of the unity of EFs into adulthood has been supported by both behavioral and neural examinations demonstrating the existence of a unifying umbrella construct termed “Common EF”^{16–19}. However, while latent variable modeling of EFs is appropriate to support the inclusion of a Common EF component at a single moment in time, it is unsuitable for testing the differentiation hypothesis and assessing the dynamic development of EFs over time.

To advance our understanding of how EFs evolve over development, we need to understand how EFs dynamically evolve as a connected network of abilities. In place of latent variable modeling, we suggest leveraging a powerful family of techniques that provides a data-driven method for understanding the number of unique cognitive mechanisms distinct from their commonalities: network analysis. Network analysis is an approach gaining traction in the psychometric field for understanding cognitive constructs comprising complex inter-related components such as intelligence, psychopathology, and personality^{20–24}. In network analysis, relationships between variables are determined after accounting for what is common among all variables. Data-driven methods can then be used to assess which variables are most closely related. In this way, network analysis allows us to examine the structure of EF using a holistic perspective to arrive at the organization that best reflects the data in a theory-agnostic, data-driven manner.

Here, we capitalize on the improved interpretability of longitudinal and cross-sectional comparisons afforded by the use of the same tasks across all participants¹⁰ with our ACE battery to shine light on the relatively understudied period of middle childhood (~7-12 years old), the developmental stage in which EFs may develop most rapidly^{25–27}. We demonstrate the benefits of network analysis to understand the organization of EFs across development by first testing the differentiation hypothesis with the historically popular latent variable analysis approach and then contrasting these results with those found using a network analysis approach. Specifically, we use each method to determine not only when components of EF become distinct from one another but, critically, when they become distinct from the unifying Common EF component. Finally, we leverage information generated from network analyses to gain insights into the stability of the organization of EFs across time. We show that during middle childhood, EF organization begins to stabilize, yet continues to develop in a manner suggesting EFs need continued support throughout their protracted development as children transition to adolescence. Developmental insights revealed by network analyses and missed by latent variable analyses may explain the inconsistencies in number of components identified across development to date and lay the groundwork for new avenues of investigation to understand how to best support EFs across the lifespan.

Results And Discussion

We first show how the use of novel, adaptive assessments can robustly measure EFs longitudinally across a wide age range without floor and ceiling effects. We then demonstrate how a holistic modeling approach can enhance our current understanding of the emergence and development of EFs by testing the differentiation hypothesis using two analytic approaches, latent variable analysis and network analysis. Using a latent variable approach, we replicate the ambiguous, difficult to interpret results found in prior investigations. We then critically extend our understanding using a network analytic approach, revealing developmental insights missed under the latent variable approach.

Novel EF Measurement

To assess EF performance with the same measures across a wide age range (7-14 years old), we used a novel mobile assessment battery, ACE (see Methods). A total of 1,280 students from three cohorts (3rd-4th grade, 5th-6th grade, and 7th-8th grade) completed ACE at four timepoints over two years (the beginning and end of the academic years 2016-2017 and 2017-2018). Here, we report data from eight tasks measuring three hypothesized EF components: working memory (WM; Forward Spatial Span, Backward Spatial Span), response inhibition (RI; Impulsive Attention, Sustained Attention, Tap and Trace) and interference resolution (IR; Stroop, Flanker, Boxed). Task grouping into these three components differs slightly from the extant literature in an effort to bring greater precision to the EFs measured. For example, while the components put forth by Miyake and colleagues²⁸ ('updating', 'inhibitory control', and 'cognitive flexibility') are the most frequently examined components¹⁵, they are often inconsistently defined across the literature. In particular, the 'inhibitory control' component is often measured with a combination of tasks that involve both the 'interference resolution' and 'response inhibition' aspects of cognitive control (see¹ for review). However, neural data from both children and adults indicate these are indeed two separate components²⁹. By separating inhibitory control-related tasks into those in which a response must always be made (interference resolution) and those in which a participant must decide whether to make a response or not (response inhibition), we are able to bring further specificity to the characterization of EFs in middle childhood.

ACE also included a measure of basic response time (BRT) and a screening assessment for red-green color blindness³⁰ at the beginning of each data collection session. Each task was developed from cognitive assessments commonly used in lab-based settings and modified for use in real-world settings by including adaptive algorithms, highly motivating trial-wise and end-of-task feedback, and a user-friendly interface (see³¹ for full details on the assessment battery and implementation methods). Importantly, the adaptive algorithms enabled two critical affordances: (1) the same tasks could be used in the same students across multiple timepoints, to reveal a student's changing cognitive abilities without being confounded by ceiling or floor effects, and (2) the same tasks could be used across students of diverse ages, to reveal individual differences in cognitive abilities across development without being confounded by the use of different tasks. This advancement in our approach to assessment enabled robust integrative data analytics within-subjects over time, and across-subjects from a wide age range.

We found each adaptive EF assessment captured predicted developmental improvements in performance. Linear mixed effects models examining task performance for each metric of interest (see Methods), allowing random effects for participant, school, and time, showed that, across tasks, performance significantly

increased with age and time after controlling for BRT, cohort, and gender (see Figure 1 and Figure S1). Beyond these predicted EF improvements with age, performance on most tasks showed a significant interaction between age and time, suggesting that younger participants tended to improve more over time compared to older participants.

To confirm the adaptive mechanism indeed challenged EFs similarly across the different grade-levels and assessment occasions, we examined in-game accuracy. In tasks with an adaptive response window (Impulsive Attention, Sustained Attention, Tap and Trace, Stroop, Flanker, and Boxed), participants only received 'correct' feedback if they provided the correct answer within a limited time frame. All other responses resulted in feedback indicating the response was 'late' or 'incorrect'. This adaptive algorithm was designed to produce ~75% of responses resulting in 'correct' feedback for all participants, and produced an average of 72.04% across tasks. While linear models examining the effect of cohort and time on percentage of trials with 'correct' feedback did show significant main effects of differences between cohort and timepoint, the effect size of these models was small (all $R^2 < 0.2$; see Table 1). Such small effects of cohort and timepoint on percentage of responses with 'correct' feedback suggests the adaptive tasks successfully presented a similar challenge across developmental stages and measurement occasions.

Table 1. Mean (Standard Deviation) In-game Accuracy for Tasks With an Adaptive Response Window and Variance Explained by Model of Cohort and Time.

	<i>3rd-4th grade cohort</i>	<i>5th-6th grade cohort</i>	<i>7th-8th grade cohort</i>	<i>Model F (NumDF, DenDF)</i>	<i>Model R²</i>	<i>95% Confidence Interval</i>
<i>Sustained Attention</i>	87.9 (6.1)	89.7 (4.2)	90.2 (5.0)	44.75 (3, 3918)	0.033	0.023, 0.045
<i>Impulsive Attention</i>	74.6 (4.8)	76.0 (4.3)	76.9 (4.1)	78.06 (3,3954)	0.056	0.043, 0.071
<i>Tap & Trace</i>	71.9 (8.9)	75.6 (8.0)	78.4(7.3)	153.38 (3, 3049)	0.131	0.111, 0.154
<i>Stroop</i>	64.0 (10.2)	62.2 (10.4)	61.2 (10.6)	304.90 (3,3800)	0.194	0.173, 0.216
<i>Flanker</i>	64.6 (7.3)	64.6 (7.1)	64.6 (7.4)	111.86 (3, 3267)	0.093	0.076, 0.113
<i>Boxed</i>	63.4 (4.3)	65.1 (3.9)	65.8 (3.9)	95.51 (3, 3755)	0.071	0.057, 0.088

Note: Data are collapsed across the four timepoints. NumDF = numerator degrees of freedom; DenDF = denominator degrees of freedom.

Novel EF Modeling

We next demonstrate that, even when solving for persistent challenges to measuring EFs, attempts to test the differentiation hypothesis using latent variable modeling yield largely uninterpretable results and provide

little insight into how to improve such models. Instead, network analysis provides a flexible, data-driven approach to understanding how the developmental process unfolds and hypothesis-generating insights into how experience might shape this development.

Latent Variable Analysis. To directly test the differentiation hypothesis using latent variable modeling, we compared a series of models to establish the number of distinguishable EF components at each stage of development using confirmatory factor analysis (CFA; see Methods, Figure S2). In accordance with the differentiation hypothesis, we expected that the number of EF components, represented by more complex models with more unique factors, would provide better model fit for older students. However, we did not explicitly incorporate Common EF into these models, as its inclusion would prevent tests statistically comparing alternate models. Instead, we examined correlations between factors to assess when these components could be differentiated beyond the unifying Common EF factor. Correlations greater than 0.7 between factors indicate that components represent redundant information (sharing more than 49% of variance) and are therefore likely not fully differentiated from one another. Based on prior adult literature and the tasks used in the current study, the number of components could range from one to three, with the maximally-differentiated organization of EFs representing WM, IR, and RI components.

Overall, despite resolving measurement constraints in prior tests of the differentiation hypothesis, the latent variable approach revealed an indeterminate developmental progression of differentiation of EF components. Generally, results suggest that a single component best describes the organization of EF from 3rd through 4th grade, after which three distinct EF components can be identified. However, this pattern is not unequivocal, and many open questions remain. Statistical comparisons (Table 2) and estimates of model fit (Table S1) indicate that for both the 5th-6th and 7th-8th grade cohort, a single component was adequate to describe the structure of EF at the first timepoint. These results suggest there may be continued refinements of EF structure over time but provide little information that can be used to begin specifying the details of this potential refinement. Further, the degree of differentiation of these factors from Common EF was unclear; factor correlations suggest WM differentiates by about 5th grade ($M_{WM \text{ and } RI} = 0.40$; $M_{WM \text{ and } IR} = 0.54$) , however, a persistent high degree of overlap between RI and IR ($M_{IR \text{ and } RI} = 0.69$) leaves open the question of whether one or both of these components would be distinguishable from Common EF (see Tables S2-S6 for full list of factor loadings and correlations). Without statistical methods to determine when components become distinct from both other EFs and Common EF, the use of latent variable models to answer questions about the differentiation hypothesis becomes even more untenable.

Table 2. *Satorra-Bentler Scaled Tests Comparing 1-, 2-, and 3-factor Models of Executive Function*

Cohort	Time-point	1- versus 2-Factor (IR with RI)		1- versus 2-Factor (WM with RI)		1- versus 2-Factor (WM with IR)		2- (IR with RI) versus 3-Factor		2- (WM with RI) versus 3-Factor		2- (WM with IR) versus 3-Factor	
		$\Delta\chi^2(\Delta df)$	<i>p</i>	$\Delta\chi^2(\Delta df)$	<i>p</i>	$\Delta\chi^2(\Delta df)$	<i>p</i>	$\Delta\chi^2(\Delta df)$	<i>p</i>	$\Delta\chi^2(\Delta df)$	<i>p</i>	$\Delta\chi^2(\Delta df)$	<i>p</i>
3rd_4th grade cohort	1	10.603 (1)	0.001	0.157 (1)	0.692	0.226 (1)	0.635	0.277 (2)	0.871	10.643 (2)	0.005	11.199 (2)	0.004
	2	4.539 (1)	0.033	28.671 (1)	< 0.001	20.036 (1)	< 0.001	21.504 (2)	< 0.001	3.486 (2)	0.175	12.052 (2)	0.002
	3	0.002 (1)	0.962	2.975 (1)	0.085	1.448 (1)	0.229	3.662 (2)	0.160	0.255 (2)	0.880	2.156 (2)	0.340
	4	2.803 (1)	0.094	3.511 (1)	0.061	11.037 (1)	0.001	10.544 (2)	0.005	10.116 (2)	0.006	2.396 (2)	0.302
5th_6th grade cohort	1	9.171 (1)	0.002	1.807 (1)	0.179	3.263 (1)	0.071	3.237 (2)	0.198	10.27 (2)	0.006	9.450 (2)	0.009
	2	9.685 (1)	0.002	18.086 (1)	< 0.001	9.551 (1)	0.002	12.839 (2)	0.002	7.828 (2)	0.020	14.254 (2)	0.001
	3	10.905 (1)	0.001	2.194 (1)	0.139	17.954 (1)	< 0.001	15.696 (2)	< 0.001	30.375 (2)	< 0.001	10.741 (2)	0.005
	4	9.229 (1)	0.002	14.337 (1)	< 0.001	38.935 (1)	< 0.001	18.364 (2)	< 0.001	13.323 (2)	0.001	7.003 (2)	0.030
7th_8th grade cohort	1	10.761 (1)	0.001	2.335 (1)	0.126	5.982 (1)	0.014	5.444 (2)	0.066	14.183 (2)	0.001	10.574 (2)	0.005
	2	13.859 (1)	< 0.001	5.111 (1)	0.024	6.898 (1)	0.009	6.740 (2)	0.034	15.111 (2)	0.001	13.203 (2)	0.001
	3	26.161 (1)	< 0.001	7.091 (1)	0.008	21.049 (1)	< 0.001	18.275 (2)	< 0.001	38.067 (2)	< 0.001	20.456 (2)	< 0.001
	4	42.235 (1)	< 0.001	20.247 (1)	< 0.001	37.352 (1)	< 0.001	33.259 (2)	< 0.001	57.775 (2)	< 0.001	41.058 (2)	< 0.001

Note. IR = interference resolution; RI = response inhibition; WM = working memory; $\Delta\chi^2$ = difference in Satorra-Bentler scaled χ^2 between nested and comparison models; Δdf = difference in degrees of freedom between nested and comparison models. Bolded values *p*-values represent cases where the more complex model shows significantly better model fit compared to the simpler model.

Moreover, alternative hypotheses around the EFs involved in different tasks are unlikely to be developed from these results. Different structures from those tested here may fit the data better (e.g., a task may index a different EF component at different developmental stages), but methods for statistically comparing such alternate hypotheses regarding which EF component a task draws on are not straightforward and would not be feasible to test without additional theoretical guidance. While there are data-driven approaches for latent variable analysis (e.g., exploratory factor analysis), these methods similarly cannot account for Common EF. As such, instead of latent variable modeling, data-driven methods that both account for what is common among EF task performance and can generate further testable hypotheses are needed to understand the complex and nuanced evolution of EF organization during middle childhood.

Network Analysis. Next, we demonstrate how using network analysis to treat EF task performances as an interconnected set of cognitive processes leads to insights into their development, which were missed using the predominant modeling approach of the field. With data-driven methods for grouping task performance according to strength of in-group performance compared to out-group performance, EF component construction was not restricted by theoretical assumptions of which tasks draw on each EF component. Further, the degree of differentiation of components identified with this method is unambiguous; components are only identified if they are distinct from the unifying Common EF component.

By applying network analysis techniques, we established a clear developmental timeline of EF organization and revealed several critical insights into how EFs evolve over time. First, while both methods (latent variable and network modeling) point to EF organization stabilizing around 5th grade, network analyses revealed that a single, undifferentiated component of EF is an unlikely organization for any age in grades 3 through 8. Second, both methods suggest great variability in the 3rd to 4th grade cohort and continued refinement from

5th to 8th grade, but network analysis reveals which EFs are developing and in what way. Finally, unlike latent variable analysis, the metrics generated from network analyses can be used to gain further insight into the development of EFs and develop new hypotheses around their trajectories. We next discuss these findings in more detail.

Both latent variable and network approaches suggest that, for younger students, there are fewer distinct EF components and that by about 5th grade, EFs can be organized into three distinct components. However, network analyses provide a clearer and more consistent developmental trajectory. Critically, the interpretation as to whether components are distinct from Common EF is straightforward in network analysis because communities are only formed if they are distinct from Common EF and from each other. We can therefore be confident that these communities represent distinct, differentiated components of EF, unlike latent variable analysis which requires interpretation of between-factor correlations. Network analysis results were also more consistent in terms of the number of differentiated EF components compared to factor analysis results. Factor analyses pointed to at least one timepoint per cohort where the number of factors that best fit the data differed from the other three timepoints. Network analysis, in contrast, only showed a different number of components across time for the 3rd-4th grade cohort (the cohort that showed the most variability in both analyses). Additionally, consistent with prior literature with this age range⁹, network models never indicated a single, undifferentiated component of EF. Further, network analysis revealed a more nuanced organization of components - task performances did not always cluster as would be predicted by theory, as described below. Using factor analysis, such differences were missed, since the model using the theory-driven organization of EFs fit the data reasonably well, and there was no indication that a different organization might be a better representation of EF constructs.

While both methods indicated the organization of EF task performances was most variable early in development, unlike latent variable modeling, network graphs (Figure 2) show that the *number* of communities for the 3rd-4th grade cohort was fairly consistent across timepoints, yet the *composition* of these communities was variable. EF organization for the older cohorts, though, was relatively stable. For both the 5th-6th grade and 7th-8th grade cohort, the tasks almost always formed three communities with groupings consistent with those predicted by theory. However, at Timepoint 2, Impulsive Attention was grouped with WM tasks for 5th-6th grade cohort, while Tap and Trace was grouped with IR tasks for the 7th-8th grade cohort. Thus, while EFs can be organized into three distinct components by about 5th grade, organization of the IR and RI components in particular continue to undergo refinement across development. See Figure S3 for estimates for all edge weights with parametric bootstrapped 95% confidence intervals.

A unique benefit of network analysis is our ability to leverage the resulting network metrics to quantify and compare the degree of network stability across cohorts. A one-way ANOVA was used to directly interrogate whether correlations between network connections (Table 3) were more variable in younger cohorts compared to older cohorts. Results revealed the within-cohort correlations indeed significantly differed across cohorts ($F(2,15) = 16.71, p < 0.001, \eta^2 = 0.69$). Tukey post-hoc tests showed correlations between the 3rd-4th grade cohort networks were significantly lower than correlations between both the 5th-6th grade cohort networks ($M_{\text{difference}} = 0.24, 95\% \text{ CI } [0.01 - 0.38], p = 0.001$) and the 7th-8th grade cohort ($M_{\text{difference}} = 0.30, 95\% \text{ CI } [0.16, 0.44], p < 0.001$). Within-cohort network correlations did not differ significantly between

these two older cohorts though ($M_{\text{difference}} = 0.05$, 95% CI [-0.08, 0.20], $p = 0.55$). Thus, the period between 3rd and 4th grade is further supported as one in which the organization of EFs is undergoing large degrees of change, whereas development between 5th and 8th grade may be more incremental. Together, these results illustrate how a holistic examination of the EF system can reveal novel insights into how these processes develop, beginning to resolve the inconsistencies across the literature that have emerged from the use of a reductionist framework.

Table 3. *Correlations of Edge Weights Between Network*

		<i>3rd - 4th grade cohort</i>				<i>5th - 6th grade cohort</i>				<i>7th - 8th grade cohort</i>			
	Timepoint	1	2	3	4	1	2	3	4	1	2	3	4
<i>3rd - 4th grade cohort</i>	1	1.00											
	2	0.37 [0, 0.65]	1.00										
	3	0.36 [-0.02, 0.64]	0.28 [-0.1, 0.59]	1.00									
	4	0.43 [0.07, 0.69]	0.31 [-0.07, 0.61]	0.44 [0.08, 0.7]	1.00								
<i>5th - 6th grade cohort</i>	1	0.56 [0.23, 0.77]	0.52 [0.18, 0.75]	0.49 [0.14, 0.73]	0.48 [0.13, 0.72]	1.00							
	2	0.32 [-0.06, 0.62]	0.28 [-0.11, 0.59]	0.38 [0.01, 0.66]	0.46 [0.11, 0.71]	0.40 [0.04, 0.68]	1.00						
	3	0.38 [0.01, 0.66]	0.44 [0.08, 0.7]	0.41 [0.04, 0.68]	0.44 [0.08, 0.7]	0.77 [0.55, 0.89]	0.67 [0.39, 0.83]	1.00					
	4	0.36 [-0.02, 0.64]	0.49 [0.14, 0.73]	0.60 [0.29, 0.8]	0.56 [0.24, 0.77]	0.60 [0.29, 0.79]	0.56 [0.24, 0.77]	0.63 [0.33, 0.81]	1.00				
<i>7th - 8th grade cohort</i>	1	0.41 [0.05, 0.68]	0.56 [0.23, 0.77]	0.41 [0.04, 0.68]	0.44 [0.08, 0.7]	0.59 [0.27, 0.79]	0.42 [0.06, 0.69]	0.64 [0.35, 0.82]	0.51 [0.17, 0.74]	1.00			
	2	0.42 [0.05, 0.68]	0.56 [0.23, 0.77]	0.47 [0.12, 0.72]	0.50 [0.15, 0.73]	0.68 [0.41, 0.84]	0.59 [0.28, 0.79]	0.72 [0.48, 0.86]	0.68 [0.42, 0.84]	0.66 [0.38, 0.83]	1.00		
	3	0.51 [0.17, 0.74]	0.68 [0.4, 0.84]	0.36 [-0.02, 0.65]	0.45 [0.09, 0.7]	0.68 [0.41, 0.84]	0.61 [0.31, 0.8]	0.77 [0.55, 0.89]	0.64 [0.36, 0.82]	0.67 [0.39, 0.83]	0.83 [0.67, 0.92]	1.00	
	4	0.51 [0.16, 0.74]	0.57 [0.25, 0.78]	0.30 [-0.08, 0.61]	0.51 [0.17, 0.74]	0.56 [0.24, 0.77]	0.60 [0.29, 0.8]	0.51 [0.17, 0.74]	0.61 [0.31, 0.8]	0.59 [0.28, 0.79]	0.64 [0.34, 0.82]	0.58 [0.26, 0.78]	1.00

Limitations

This study makes significant strides in our approach to measure and model EFs, improving on several critical limitations in the field. Yet, further advancements must continue to be made to build upon and address limitations of the work presented here, particularly in regard to the scope of EFs assessed and the availability of statistical methodologies to compare network models longitudinally.

Developing a novel, adaptive battery of EF tasks for all ages and abilities was not without its challenges; a design decision resulted in limitations to our ability to robustly measure all EF components. Specifically, the WM component was only indexed by two measures, which limited the type of latent variable model that could be constructed and tested here. While a third task hypothesized to measure WM, Filter, was originally included in the ACE battery, it required a different adaptive mechanism, which resulted in age-related differences in challenge level, leading to its exclusion from the current analysis. Consequently, we could not test certain factor configurations (see Methods) without rendering the models uninformative.

Further, as discussed previously, the list of EFs measured here is not exhaustive. In particular, the components ‘updating’ and ‘cognitive flexibility’ as popularized by Miyake and colleagues’ work modeling EFs in adults²⁸ were not assessed here. Due to time constraints associated with in-school testing sessions, we were limited to the ten tasks discussed above, thus we chose to focus on particularly understudied aspects of EF. Using the adaptive assessment framework we outlined here, it is our intent that additional EF assessments tapping other components are developed. Future work can then incorporate these additional components in the network analytic framework to expand the developmental model introduced here.

There were also methodological challenges related to comparing two analytical approaches (latent variable and network analyses). We intentionally did not explicitly model the dependency of multiple observations per student that occurs with longitudinal data in either analytic approach. While it is possible to model this dependency using factor analysis, development of network models that can handle longitudinal data are still in their infancy (though see³²). To keep the general modeling strategy consistent so that inferences would be comparable, we treated all observations as independent in both approaches. However, this strategy is unlikely to have affected the results for two reasons. First, without accounting for within-person changes, within-cohort comparisons were more conservative than necessary. Second, we did not perform tests that may have been more likely to be affected by treating observations as independent, such as testing for differences in the strengths of within-cohort associations between task performance. Nonetheless, as network analytic methodology continues to advance, it will be important to also advance the methods used to reveal the evolution of EF structure across development.

Finally, neither modeling approach was able to simultaneously account for Common EF *and* provide statistical comparisons between models of differing complexity. With latent variable analysis, it is a straightforward process to compare whether a model with more factors fits statistically better than a model with fewer factors. These capabilities, though, are currently limited with network models (though see³³). Community detection algorithms provide a likely grouping for task performances, but there is no index to determine whether a two-community network explains EFs just as well as a three-community network, for example. However, existing methods for accounting for Common EF in the latent variable approach preclude such statistical comparisons between models, leaving the theoretical problem of how to account for

Common EF in the context of differentiation of components with this approach unresolved. To date, the benefits of the network analysis approach, which accounts for commonality among all EF task performances rather than treating it as a separate component entirely, presents a promising solution for accounting for Common EF. The rapidly emerging statistical approaches for testing network model complexity position this technique as the path forward in establishing the developmental trajectory of EFs.

Conclusions And Future Directions

This study presents potentially game-changing methods for understanding precisely how EFs differentiate across middle childhood. Adaptive algorithms in our EF assessments allow us to meet the learner where they are, regardless of ability, and allow for multiple assessments within-subject over time. The network analytic approach offers exciting new avenues for understanding the development of EF as a dynamic interconnected network of skills that can align behavioral and neural models. With more methodological advancements on the horizon to improve the statistical precision of network modeling results, we expect this approach will critically advance our understanding of how a learner's EF structure evolves across middle childhood and beyond.

The future potential for network analysis to understand complex cognitive constructs is bright. Researchers in related fields have already begun to capitalize on information gained from taking a network analytic perspective to understand other cognitive processes. For example, Kan and colleagues³⁴ demonstrated how fit statistics can be obtained for network models, allowing a direct comparison between network and latent variable models. As such, future research could directly compare a variety of configurations of EF modeled using latent variable analysis to those using network analysis to determine which organization best fits observed EF performance. While outside the scope of the current paper, researchers in the field of intelligence have used this approach to show that modeling aspects of intelligence as being mutually and reciprocally related through a network framework is favored over modeling an overarching umbrella component ('g') in a latent variable framework³⁵. Given this field's similar dilemma around how to quantify developmental differentiation in the presence of task commonality³⁶, we anticipate such investigations in EFs will be similarly fruitful for elucidating the mechanisms through which skill changes arise.

Further, as methods for appropriately modeling longitudinal data emerge, network analysis provides an avenue for understanding the potential reciprocal relationships among EFs over time³². For example, in a separate study we are examining how growth in performance on individual tasks are connected. By using a network framework for investigating EF skill growth, we can evaluate whether the same communities formed when modeling contemporaneous ties between task performances also emerge when looking at their patterns of growth across time. Such evidence would reinforce the identity of the communities as distinct components of EF and allow us to answer whether components of EF emerge independently or in tandem with other components.

Such insights into the development of EFs are critical for advancing our understanding of how they influence, and can be influenced by, internal and external factors. For example, EFs are often the focus of educational interventions with the goal of improving academic-related outcomes (see e.g.,³⁷⁻³⁹). Network

analysis is well-poised to generate hypotheses regarding which EF tasks or components might be more likely to transfer outside a training regime, which can then guide future training studies. Indeed, the findings from the current study provide a clear set of testable hypotheses: given that the cross-sectional network models found here suggest that WM is less strongly connected to other EF components, future training studies should test the hypothesis that training a highly connected component such as IR would be more likely to result in transfer to other EFs compared to training on the less-well connected WM component.

The findings from this study showcase how advances in assessing EFs and an increasingly popular modeling technique, network analysis, can be applied to the field of EFs to better align behavioral and neural investigations. The dual paradigm shifts to network analysis using adaptive measures provide a promising pathway for refining and specifying our understanding of how EFs develop. These insights can in turn be applied to advance our understanding of EFs' wide-reaching impact on factors related to physical and cognitive health across the lifespan². Together, our improved methodological approaches to *measuring* EFs can lead to the development of improved methods for *training* EFs and providing students the proper foundation they need for learning and future educational success.

Methods

Participants in the current study were recruited through their schools as part of Project iLEAD (in-school longitudinal executive function and academic achievement database), a two-year accelerated longitudinal study of EF development students grades 3-8. Full details of Project iLEAD are reported in³¹ and summarized here.

The study was approved by the Institutional Review Board (IRB) of the University of California, San Francisco and was conducted in accordance the relevant guidelines and regulations. Written parental or guardian consent was obtained from all participants at the beginning of the study, and verbal assent from all participants was obtained before all in-class data collection sessions. At the end of the study, all students in participating classrooms received snacks and stickers, regardless of participation.

Participants

Nine schools (seven public, one independent, one parochial) from northern California opted to participate in this longitudinal study, which included assessments at the Fall and Spring of two academic years for a total of four assessment periods. In total, 1,280 students participated over the course of two years. At the beginning of each school year, teachers distributed consent forms to students to take home for parental or guardian review and signature. This first round of recruitment resulted in a total of 1,088 participating students in Year 1: 284 third graders (mean age 8.07 years old, $SD = 0.35$), 260 fifth graders (mean age 9.98 years old, $SD = 0.41$), and 544 seventh graders (mean age 11.9 years old, $SD = 0.47$). In the fall of Year 2, we re-opened enrollment to participating classrooms to allow new students to participate in the study, which resulted in an additional 195 students joining the study (44 fourth, 147 sixth, and 4 eighth grade students). The Year 2 sample thus included 1,106 students: 288 fourth graders (mean age of 9.03 years old, $SD = 0.33$), 336 sixth graders (mean age of 10.9 years old, $SD = 0.39$), and 482 eighth graders (mean age of 12.9 years old, $SD = 0.44$).

Our sample was demographically diverse; see Table 4 for demographics of participating students at each of the four timepoints. Gender category reflected the self-identified gender of the students as expressed to the district and researchers (separately). Additional demographic data (see below) were provided for students enrolled in public schools and whose parents consented to share this information ($n = 1,159$ of 1,280).

Table 4. *Student Demographics*

		<i>Timepoint 1</i>		<i>Timepoint 2</i>		<i>Timepoint 3</i>		<i>Timepoint 4</i>	
		n	%	n	%	n	%	n	%
<i>Gender</i>	Male	519	50.9	514	51.8	541	51.3	510	51.2
	Female	500	49.1	478	48.2	513	48.7	487	48.8
<i>Ethnicity</i>	American Indian or Native Alaskan	5	0.5	6	0.6	6	0.6	5	0.5
	Asian	339	33.3	332	33.5	369	35.0	350	35.1
	Black or African American	20	2.0	18	1.8	18	1.7	17	1.7
	Filipino	55	5.4	56	5.7	60	5.7	56	5.6
	Hispanic or Latinx	267	26.2	253	25.5	280	26.6	269	27.0
	Pacific Islander	7	0.7	6	0.6	5	0.5	6	0.6
	Two or more ethnicities	43	4.2	45	4.5	44	4.2	44	4.4
	White	170	16.7	165	16.6	185	17.6	169	17.0
	Unknown	113	11.1	111	11.2	87	8.26	81	8.1
<i>Special Education Status</i>	No	806	79.1	771	77.7	855	81.1	807	80.9
	Yes	101	9.9	111	11.2	113	10.7	111	11.1
	Unknown	112	11.0	110	11.1	86	8.2	79	7.9
<i>Low Income Qualification</i>	No	583	57.2	570	57.5	616	58.4	587	58.9
	Yes	324	31.8	312	31.5	352	33.4	331	33.2
	Unknown	112	11.0	110	11.1	86	8.2	79	7.9
<i>English Language Fluency</i>	English monolingual	396	38.9	391	39.4	434	41.2	413	41.4
	English multilingual, never enrolled in English classes	56	5.5	58	5.9	67	6.4	63	6.3
	English multilingual, previously enrolled in English classes	308	38.9	289	29.1	324	30.7	311	31.2
	Current English Language Learner	147	14.4	144	14.5	143	13.6	131	13.1
	Unknown	112	11.0	110	11.1	86	8.2	79	7.9

Procedures

We administered a series of mobile assessments of EF, math, and reading skills that took the form of digital ‘games’, during school hours, at the beginning and end of each academic year (fall and spring) over two school years (see Assessment Administration details below). At each of the four timepoints, EF assessments were administered during one class period, with the research team returning a little over a month later to administer the math and reading assessments ($M = 5.7$ weeks, $SD = 2.4$, min. = 1.9, max. = 10). At the end of each academic year, academic performance and other relevant data were provided by the district for students whose parents consented to share district data.

All tasks were administered in a group setting on iPads. Administration was facilitated by a large team of research staff, in groups that ranged from seven to 83 students ($M = 30$). A lead facilitator gave verbal instructions to the group for each task, aided by visual instructions from a 24" x 36" flipbook. Participants began each task at the same time, and instructions for the next task were not given until all participants completed the current task. Each task began with practice trials during which researchers monitored participants to ensure participants understood the task and were correctly following task instructions. Researchers closely monitored the sessions throughout administration to provide technical assistance if necessary, answer student questions, and monitor performance. Administration sessions lasted approximately 50 minutes.

Assessments

Adaptive Cognitive Evaluation (ACE). This study used a novel mobile assessment battery, *ACE*, to assess EF skills. *ACE* was developed from cognitive assessments commonly used in lab-based settings and modified for use in real-world settings by including adaptive, psychometric staircase algorithms, and trial-wise as well as end-of-task feedback¹². For this study, we additionally implemented a user-friendly interface and task tutorials for students in middle childhood. Importantly, the psychometric staircase algorithms allowed for the same tasks to be used in the same students over time, which enables benchmarking of an individual’s changing cognitive abilities without being confounded by ceiling or floor. Further, these adaptive algorithms allowed for the comparison of the same tasks across students of different ages, genders, races, and cultures¹². The assessment battery consisted of a color blindness test, a response time control task, two working memory tasks, an attentional filtering task, two response inhibition tasks, three interference resolution tasks, and a cognitive flexibility task. The attentional filtering and cognitive flexibility task were excluded from the current analysis due to differential task challenge across grade levels (attentional filtering task) or technical errors that prevented consistent data reporting across timepoints (cognitive flexibility; see³¹ for more details). All other tasks are briefly described below, and example stimuli and schematics for each task are presented in Figure S4.

Response Time Control Task. The first *ACE* task was a measure of basic response time (*BRT*; Figure S4A). Because improvements in EFs have also been associated with improvements in general processing speed (e.g.,^{40,41}), *BRT* was designed to serve as a covariate to be regressed from performance metrics of all other *ACE* tasks.

Color Blindness Test. The second *ACE* task was a screening assessment for red-green color blindness³⁰ (Figure S4B).

Working Memory. Two tasks were used to measure working memory, *Forward Spatial Span* (Figure S4C) and *Backward Spatial Span* (Figure S4D). These two tasks were digital modifications based on the Corsi Block Task⁴². In this task, students were shown an array of open circles, with a target sequence cued via circles becoming filled, in sequence, with either green (*Forward Spatial Span*) or blue (*Backward Spatial Span*) color. Sequence length increased according to performance (see Adaptivity below). Once students viewed the cued sequence, they were instructed to recreate the sequence in the same order (*Forward Spatial Span*) or in the reverse order (*Backward Spatial Span*).

Response Inhibition. Response inhibition was measured with three tasks, Sustained Attention, Impulsive Attention (collectively referred to as the *Continuous Performance Task (CPT)*; Figure S4F) and *Tap and Trace* (Figure S4G). For both tasks, students were instructed to respond to a target stimulus and withhold a response to non-target stimuli. *CPT* is a target detection task adapted from the Test of Variable Attention⁴³ (TOVA). This task included two blocked conditions: a target frequent condition (80% target trials) designed to assess impulse control (*Impulsive Attention*) and a target infrequent condition (20% target trials) designed to test sustained attention abilities (*Sustained Attention*). *Tap and Trace* is a dual-task assessment adapted from the paradigm described by Eversheim and Bock⁴⁴. This task included three blocked conditions: one in which students used their dominant hand to tap when they detected a target stimulus, a second in which they traced a shape with their non-dominant hand, and a third in which they performed both tasks simultaneously.

Interference Resolution. Interference resolution was measured with three tasks: *Stroop* (Figure S4H), *Flanker* (Figure S4I), and *Boxed* (Figure S4J). *Stroop* is based on the computerized version of the color-word *Stroop* task as described by Mead and colleagues⁴⁵ in which students selected the text color (e.g., green) of a centrally presented color word (e.g., BLUE). On 30% trials, the text and word were congruent, and on 70% trials they were incongruent. *Flanker* is a letter flanker task based on the paradigm described by Eriksen and Eriksen⁴⁶ in which students are instructed to indicate the middle letter of a string of five letters. On 50% trials, the central and flanking letters were congruent, and on 50% trials they were incongruent. Finally, *Boxed* is a top-down/bottom-up attention task based on the visual search paradigm first described by Treisman and Gelade⁴⁷ in which students must identify a target stimulus in an array of distractor stimuli. This task included four blocked conditions that varied on search condition and number of distractors. In each condition, the target was either identifiable by one feature (color) or by the conjunction of two features (color of target and location of opening of the target box) and either a low (3) or high (11) number of distractor stimuli.

Analysis Methods

For each *ACE* task, we *a priori* defined a metric that would best reflect the construct of interest. It is important to note that the psychometrics of each assessment dictated the *a priori* defined metrics of interest. We describe the rationale to assess each construct of interest in turn below. For *BRT*, the metric of interest was

mean response time collapsed across both dominant and nondominant hands. For the *Color Blindness Test*, we assessed whether students selected a response indicating red-green color blindness according to scoring guidelines in Ishihara (1972). The metric of interest for *Forward* and *Backward Spatial Span* was span length, or the maximum number of spatial locations in the correct sequence held in mind. For *Sustained Attention*, we used a metric that is sensitive to lapses in attention—the standard deviation of the response time to infrequently presented targets⁴⁸. For *Impulsive Attention*, we used a metric that would measure detection of targets while accounting for withholding prepotent responses to frequent non-targets, the signal detection metric of d' ⁴⁸. This metric is an index of sensitivity reflecting the ability of a student to faithfully detect a target as measured by the difference between the normalized hit rate and the normalized false-alarm rate⁴⁹. For the dual-task assessment (Tap and Trace), the metric of interest was how reliably students could detect a target vs. a distractor during the dual-task portion of the task; thus, we again deployed d' . For tasks in which students were expected to respond to each trial, we used Rate Correct Score (RCS) to index performance on both response time and accuracy. RCS was thus used for *Stroop* and *Flanker*. Task-level RCS was calculated by dividing the number of correct responses by the product of mean response time for all trials and the total number of trials responded to^{50,51}. For the visual search task, *Boxed*, a technical error resulted in trials in which students responded outside the response window to be considered incorrect (rather than correct and late). Because correct but late trials could not be distinguished from incorrect trials, RCS could not be calculated in the same manner as the other tasks. Instead, we used mean response time to all correct trials as the metric of interest for this task.

Data cleaning procedures. A very small number of students who were red-green colorblind as indicated by the colorblind screener ($n = 16$) were excluded from analysis, given that several tasks required students to discriminate between targets and distractors based on color. Trials where no response was given when a response was expected or that were anticipatory (response time < 200ms) were excluded from analyses (1.8% of all trials). All remaining trials were evaluated for accuracy regardless of whether or not the response was within the response window. Thus, trials were only considered incorrect if an incorrect response was made (and not if they were correct but occurred outside the response window).

Data from each student were evaluated and cleaned on a task-level basis at each timepoint. For each task, to be included in data analysis, students must have answered a minimum of five trials per condition and achieved above-chance accuracy on the easiest condition (i.e., the condition that required lowest cognitive load; see³¹ for a full description of task conditions and how chance-level performance was determined for each task). Data from each task were then evaluated for outlier students based on performance within each cohort and timepoint. Outlier performance was defined as performance falling outside three median absolute deviations (MADs) of the median performance of the relevant cohort at a given timepoint⁵². Finally, additional outlier analyses to identify influential observations in the larger regression analysis of task performance were conducted by computing Cook's distance. Observations with Cook's $d > 1$ were removed⁵³. These cleaning procedures resulted in exclusion of 1.9% of task-level data collected.

Confirmatory Factor Analysis. While data-driven organizations of variables are possible within an exploratory factor analysis (EFA) framework, exploratory approaches do not provide a straightforward way to account for the high degree of overlap between performance on EF tasks, and assignment of a behavior to a latent

variable is dubious, often resulting in uninterpretable organizations⁵⁴. Thus, we conducted separate confirmatory factor analyses (CFAs) for the three cohorts at the four timepoints to avoid assuming the structure of EF remained the same across all timepoints for any group, and to assess the stability of these structures over a two-year measurement period. We evaluated five models of EF - the maximally differentiated structure with three distinct factors, all possible permutations of a two-factor model in which two of the three factors are collapsed into one, and the simplest structure in which all tasks represent a single, undifferentiated EF factor (see Figure S2).

All CFAs were conducted in Mplus version 8.1⁵⁵ with the robust maximum likelihood estimation method. To statistically compare nested models, we used Satorra-Bentler scaled chi-square tests with degrees of freedom equal to the difference in number of free parameters between the comparison and nested models⁵⁶. These tests help us to determine whether more complex representations of EF are a better fit to EF task performance across middle childhood. Because these statistics are meant to compare nested models, the 1-factor model was compared to each of the 2-factor models, and each of the 2-factor models were compared to the 3-factor model, but the 2-factor models cannot be statistically compared to each other in this manner. In interpreting these results, we took a conservative approach in which a more complex model would be selected over a less complex model only if a more complex model would always be preferred, regardless of which 2-factor permutation was considered. The results of chi-square difference testing were corroborated by converging evidence from the Comparative Fit Index (CFI), root mean square error of approximation (RMSEA), Akaike Information Criteria (AIC), and sample-size adjusted Bayesian Information Criterion (BICc). CFI values > 0.90 were considered excellent model fit, with values closer to 1 indicating better model fit. RMSEA values less than or equal to .06 were considered adequate model fit⁵⁷, with lower values indicative of better model fit.

Models explicitly incorporating a Common EF factor were not tested here, as models in which Common EF is a higher-order factor do not allow us to test the differentiation hypothesis. While Common EF can be incorporated as a higher-order umbrella component reflecting what is common among all lower-order components, such an approach is not amenable to testing the differentiation hypothesis. The earliest stages of differentiation in which less than three components are differentiated from Common EF would not provide meaningful insight into the patterns of behaviors being modeled as they would not be considered properly identified⁵⁸. An alternative approach is to incorporate Common EF as an additional lower-order latent variable, rather than a hierarchical superordinate variable. In such 'bifactor' models, each observed variable measures two latent variables: Common EF and another differentiated component. Although these models can be easier to identify in some instances, it can be difficult for such complex models to converge given the historically low power and task reliability observed in extant examinations of EF¹⁵. While the hypothesis could be tested if Common EF were incorporated as a bifactor model, such a model would not be identified for this dataset without assuming performance on the WM tasks contributes equally to both the WM and Common EF factors, which has not been supported in the literature^{16,59}.

Network Analysis. Replicating the general approach used for the latent variable models, we created separate models of EF performance for each cohort and timepoint. All network analyses were conducted in R 4.0.4⁶⁰. Network models were estimated using the bootnet package^{61,62}. All models were fully saturated

partial correlation networks (non-regularized Gaussian Markov random fields), and missing data were handled via full information maximum likelihood. After estimating each network model, the Spinglass algorithm⁶³ from the igraph package^{64,65} was applied separately to each network to determine communities of tasks. Here, the Spinglass algorithm was selected over other community detection algorithms such as the Louvain algorithm because it is able to handle negative partial correlations in a network. To ensure the stability of groupings, community detection was performed 1,000 times and the most frequent grouping is reported here. Resulting network and community detection results were displayed graphically using the qqgraph package^{66,67}. For graphing purposes, nodes were fixed to the same positions across networks and partial correlations between -0.1 and 0.1 are not displayed. To understand network stability over time, edge weights from each network were correlated with each other. Because these edge weights represent partial correlations, edge weights were first Fischer transformed before being correlated with edge weights from each other network.

References

1. Diamond, A. Executive Functions. *Annu. Rev. Psychol.* **64**, 135–168 (2013).
2. Zelazo, P. D., Blair, C. B. & Willoughby, M. T. Executive Function: Implications for education. 1–148 (2016).
3. Morra, S., Panesi, S., Traverso, L. & Usai, M. C. Which tasks measure what? Reflections on executive function development and a commentary on Podjarny, Kamawar, and Andrews (2017). *J. Exp. Child Psychol.* **167**, 246–258 (2018).
4. Niendam, T. A. *et al.* Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. *Cogn. Affect. Behav. Neurosci.* **12**, 241–268 (2012).
5. Doebel, S. Rethinking executive function and its development. *Perspect. Psychol. Sci.* **15**, 942–956 (2020).
6. Mungas, D. *et al.* NIH toolbox cognition battery (CB): Factor structure for 3 to 15 year olds. *Monogr. Soc. Res. Child Dev.* **78**, 103–118 (2013).
7. Shing, Y. L., Lindenberger, U., Diamond, A., Li, S. C. & Davidson, M. C. Memory maintenance and inhibitory control differentiate from early childhood to adolescence. *Dev. Neuropsychol.* **35**, 679–697 (2010).
8. Johnson, M. H. Interactive Specialization: A domain-general framework for human functional brain development? *Dev. Cogn. Neurosci.* **1**, 7–21 (2011).
9. Lee, K., Bull, R. & Ho, R. M. H. Developmental changes in executive functioning. *Child Dev.* **84**, 1933–1953 (2013).
10. Best, J. R. & Miller, P. H. A developmental perspective on executive function. *Child Dev.* **81**, 1641–1660 (2010).

11. Anguera, J. A. *et al.* Video game training enhances cognitive control in older adults. *Nature* **501**, 97–101 (2013).
12. Anguera, J. A., Jordan, J. T., Castaneda, D., Gazzaley, A. & Areán, P. A. Conducting a fully mobile and randomised clinical trial for depression: Access, engagement and expense. *BMJ Innov.* **2**, 14–21 (2016).
13. Anguera, J. A. *et al.* Characterizing cognitive control abilities in children with 16p11.2 deletion using adaptive ‘video game’ technology: A pilot study. *Transl. Psychiatry* **6**, (2016).
14. Miyake, A., Emerson, M. J. & Friedman, N. P. Assessment of executive functions in clinical settings: Problems and recommendations. *Semin. Speech Lang.* **Volume 21**, 0169–0183 (2000).
15. Karr, J. E. *et al.* The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychol. Bull.* **144**, 1147–1185 (2018).
16. Friedman, N. P. *et al.* Individual differences in executive functions are almost entirely genetic in origin. *J. Exp. Psychol. Gen.* **137**, 201–225 (2008).
17. Friedman, N. P. & Miyake, A. Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex* **86**, 186–204 (2017).
18. Smolker, H. R., Friedman, N. P., Hewitt, J. K. & Banich, M. T. Neuroanatomical correlates of the unity and diversity model of executive function in young adults. *Front. Hum. Neurosci.* **12**, 283 (2018).
19. Reineberg, A. E., Andrews-Hanna, J. R., Depue, B. E., Friedman, N. P. & Banich, M. T. Resting-state networks predict individual differences in common and specific aspects of executive function. *Neuroimage* **104**, 69–78 (2015).
20. Borsboom, D. & Cramer, A. O. J. Network analysis: An integrative approach to the structure of psychopathology. *Annu. Rev. Clin. Psychol.* **9**, 91–121 (2013).
21. Fried, E. I. *et al.* Mental disorders as networks of problems: a review of recent insights. *Social Psychiatry and Psychiatric Epidemiology* **52**, (2017).
22. Golino, H. F. & Demetriou, A. Estimating the dimensionality of intelligence like data using Exploratory Graph Analysis. *Intelligence* **62**, 54–70 (2017).
23. van der Maas, H. L. J., Kan, K. J., Marsman, M. & Stevenson, C. E. Network models for cognitive development and intelligence. *J. Intell.* **5**, 1–17 (2017).
24. Costantini, G. *et al.* State of the aRt personality research: A tutorial on network analysis of personality data in R. *J. Res. Pers.* **54**, 13–29 (2015).
25. Anderson, P. Assessment and development of executive function (EF) during childhood. *Child Neuropsychol.* **8**, 71–82 (2002).

26. Best, J. R., Miller, P. H. & Naglieri, J. A. Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learn. Individ. Differ.* **21**, 327–336 (2011).
27. Romine, C. B. & Reynolds, C. R. A model of the development of frontal lobe functioning: Findings from a meta-analysis. *Applied Neuropsychology* **12**, 190–201 (2005).
28. Miyake, A. *et al.* The unity and diversity of executive functions and their contributions to complex “Frontal Lobe” tasks: A latent variable analysis. *Cogn. Psychol.* **41**, 49–100 (2000).
29. Bunge, S. A., Dudukovic, N. M., Thomason, M. E., Vaidya, C. J. & Gabrieli, J. D. E. Immature frontal lobe contributions to cognitive control in children: Evidence from fMRI. *Neuron* **33**, 301–311 (2002).
30. Ishihara, S. *Tests for colour-blindness*. (Kanehara Shuppan Co., 1972).
31. Younger, J. W. *et al.* Development of executive function in middle childhood: a large-scale, in-school, longitudinal investigation. Pre-print at <https://psyarxiv.com/xf489/> (2021)
32. Deserno, M., Sachisthal, M. & Epskamp, S. A magnifying glass for the study of coupled developmental changes: Combining psychological networks and latent growth models. Pre-print at <https://psyarxiv.com/ngfxq/> (2021).
33. Epskamp, S., Isvoranu, A.-M. & Cheung, M. W.-L. Meta-analytic Gaussian Network Aggregation. *Psychom.* **2021** 1–35 (2021).
34. Kan, K. J., De Jonge, H., Van Der Maas, H. L. J., Levine, S. Z. & Epskamp, S. How to compare psychometric factor and network models. *J. Intell.* **8**, 1–10 (2020).
35. Kan, K. J., van der Maas, H. L. J. & Levine, S. Z. Extending psychometric network analysis: Empirical evidence against g in favor of mutualism? *Intelligence* **73**, 52–62 (2019).
36. Molenaar, D., Dolan, C. V., Wicherts, J. M. & van der Maas, H. L. J. Modeling differentiation of cognitive abilities within the higher-order factor model using moderated factor analysis. *Intelligence* **38**, 611–624 (2010).
37. Diamond, A. & Lee, K. Interventions shown to aid executive function development in children 4 to 12 years old. *Science* **333**, 959–64 (2011).
38. Jacob, R. & Parkinson, J. The potential for school-based interventions that target executive function to improve academic achievement: A review. *Rev. Educ. Res.* **85**, 512–552 (2015).
39. Titz, C. & Karbach, J. Working memory and executive functions: effects of training on academic achievement. *Psychological Research* **78**, 852–868 (2014).
40. Christ, S. E., White, D. A., Mandernach, T. & Keys, B. A. Inhibitory control across the life span. *Dev. Neuropsychol.* **20**, 653–669 (2001).

41. Fry, A. F. & Hale, S. Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychol. Sci.* **7**, 237–241 (1996).
42. Corsi, P. M. Human memory and the medial temporal region of the brain. *Diss. Abstr. Int.* **34**, 891 (1973).
43. Greenberg, L., Lark, R., Dupuy, T. & Corman, C. The Test of Variables of Attention (TOVA). (1991).
44. Eversheim, U. & Bock, O. Evidence for processing stages in skill acquisition: A dual-task study. *Learn. Mem.* **8**, 183–189 (2001).
45. Mead, L. A. *et al.* Neural basis of the Stroop interference task: Response competition or selective attention? *J. Int. Neuropsychol. Soc.* **8**, 735–742 (2002).
46. Eriksen, B. & Eriksen, C. W. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Percept. Psychophys.* **16**, 143–149 (1974).
47. Treisman, A. M. & Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **12**, 97–136 (1980).
48. Lark, R. A., Greenberg, L. M., Kindschi, C. L., Dupuy, T. R. & Hughes, S. J. *The TOVA Professional Manual*. (The Tova Company, 2018).
49. Green, D. M. & Swets, J. A. *Signal detection theory and psychophysics*. (Wiley & Sons, Inc., 1966).
50. Vandierendonck, A. A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behav. Res. Methods* **49**, 653–673 (2017).
51. Woltz, D. J. & Was, C. A. Availability of related long-term memory during and after attention focus in working memory. *Mem. Cognit.* **34**, 668–684 (2006).
52. Leys, C., Ley, C., Klein, O., Bernard, P. & Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* **49**, 764–766 (2013).
53. Cook, R. D. Detection of influential observation in linear regression. *Technometrics* **19**, 15–18 (1977).
54. Brocki, K. C. & Bohlin, G. Executive functions in children aged 6 to 13: A dimensional and developmental study. *Dev. Neuropsychol.* **26**, 571–593 (2004).
55. Muthén, L. K. & Muthén, B. O. User 's Guide Manual. 1–13 (2017).
56. Satorra, A. & Bentler, P. M. Ensuring positiveness of the scaled difference Chi-square test statistic. *Psychometrika* **75**, 243–248 (2010).
57. Hu, L. T. & Bentler, P. M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Model.* **6**, 1–55 (1999).

58. Kline, R. B. Principles and practice of structural equation modeling, 3rd ed. *Principles and practice of structural equation modeling, 3rd ed.* xvi, 427–xvi, 427 (2011).
59. Friedman, N. P., Miyake, A., Robinson, J. L. & Hewitt, J. K. Developmental trajectories in toddlers' self-restraint predict individual differences in executive functions 14 years later: a behavioral genetic analysis. *Dev. Psychol.* **47**, 1410–30 (2011).
60. Core Development Team, R. A language and environment for statistical computing. *R Found. Stat. Comput.* **2**, <https://www.R-project.org> (2020).
61. Epskamp, S., Borsboom, D. & Fried, E. I. Estimating psychological networks and their accuracy: A tutorial paper. *Behav. Res. Methods* **50**, 195–212 (2018).
62. Epskamp, S. bootnet: Bootstrap methods for various network estimation routines. *R-package* (2015).
63. Reichardt, J. & Bornholdt, S. Statistical mechanics of community detection. *Phys. Rev. E* **74**, 016110 (2006).
64. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst. Complex Sy.* **1695** (2006).
65. Csardi, M. G. Package 'igraph': Network Analysis and Visualization. (2015).
66. Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D. & Borsboom, D. Qgraph: Network visualizations of relationships in psychometric data. *J. Stat. Softw.* **48**, (2012).
67. Epskamp, S., Costantini, G., ... J. H.-R. & 2021, U. Package 'qgraph'. *R-package* (2017).

Declarations

Acknowledgements

This research was supported by funding from the National Science Foundation, Science of Learning Collaborative Networks Grant (NSFSLCN-1540854) awarded to Melina Uncapher (lead Principal Investigator [PI]) and Adam Gazzaley, and co-PIs Joaquin Anguera, Silvia Bunge, Fumiko Hoeft, Bruce McCandliss, Jyoti Mishra, and Miriam Rosenberg-Lee. The study was performed in accordance with protocols approved by the Institutional Review Board (IRB) of the University of California San Francisco. The authors would like to thank the research staff, Jordin Rodondi, Caleb Banks, Zoe D'Esposito, John David Lorentz and the large team of UCSF volunteers as well as the students, teachers, parents, and school and district administrators who made this research program possible. The authors are also grateful for consultation from Stephanie Haft, Ariel Starr, Joshua Jordan, and Britte Cheng. We are also grateful to the team of developers who have made the ACE assessment possible including WoWLabz, Zynga.org, and Rose Feldman, and the programmers that created the aceR processing code Jose Gallegos and Monica Thieu.

Author Contributions

J.A.A., S.A.B., F.H., B.D.M., J.M., M.R.-L., A.G., and M.R.U. conceived of and designed the study. J.W.Y. and M.R.U. collected the data. J.W.Y., K.D.O., E.E.F., and M.R.U. analyzed the data. J.W.Y., K.D.O., and M.R.U. wrote the paper. All authors discussed the results and contributed to editing the manuscript.

Competing Interests

The authors declare no competing interests.

Figures

Figure 1

Growth in performance on executive function metrics of interest for each task and cohort. With few exceptions, all participants improved over time, and younger students tended to show the most gains over time, as indicated by significant main effects. Shaded region represents 95% confidence interval of linear regression of time on performance.

Figure 2

Results of network analysis and community detection for each cohort and timepoint. Edges between -0.1 and 0.1 are not shown for visualization purposes. Community detection algorithms indicate a two-community organization for the 3rd – 4th grade cohort that differentiates into a three community structure by about 5th grade. Fluctuations in grouping and magnitude of edge weights across older cohorts suggests continued subtle development for older students. WM = working memory; RI = response inhibition; IR = interference resolution; B. Span = Backward Spatial Span; F. Span = Forward Spatial Span; Sust Attn = Sustained Attention; Impul Attn = Impulsive Attention.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ProjectiLEADSupplemental.docx](#)