Supplemental Results

Longitudinal changes in block and event-related fMRI task performance during childhood

We analyzed behavioral data from the longitudinal block (N = 28) and event-related (N = 20) fMRI experiments separately. We first conducted a 2-by-2 analysis of variance (ANOVA) with Time (Time-1 vs. Time-2) and Condition (Addition vs. Control) as within-subject factors to analyze longitudinal changes in accuracy data from 28 children in the block fMRI task. This analysis revealed significant main effect of Condition (F(1, 27) = 13.04, P < 0.001), but no main effect of Time (F(1, 27) = 0.51, P = 0.48). Follow-up tests revealed generally lower accuracy for solving Addition than Control problems (Time-1: (t(27) = -2.74, P = 0.01; Time-2, t(27) = -3.56, P = 0.001)(**Fig. S1a**). There is no significant interaction between Time and Condition (F(1, 27) = 0.38, P = 0.54) for accuracy in the block fMRI experiment. We then conducted another 2-by-2-ANOVA for reaction times (RTs) for the block fMRI task. This analysis revealed significant main effects of Condition (F(1, 27) = 46.67, P < 0.001) and Time (F(1, 27) = 14.82, P = 0.001). Follow-up tests revealed that children showed generally slower response in solving Addition than Control problems (Time-1: t(27) = -4.64, P < 0.001); Time-2: t(27) = -6.10, P < 0.001), and they became faster at Time-2 than Time-1 in solving problems (Addition: (t(27) = 2.65, P = 0.013; Control: t(27) = 4.45, P < 0.001) (**Fig. S1c**). We did not observe significant interaction effect between Condition and Time (F(1,27) = 0.90, P = 0.35).

For the longitudinal event-related fMRI experiment, one 2-by-2 ANOVA for accuracy data revealed significant main effect of Condition (F(1, 19) = 46.24, P < 0.001) and marginally significant main effect of Time (F(1, 19) = 3.07, P = 0.096). Fellow-up tests revealed that children showed generally lower accuracy in solving Addition than Control problems (Time-1: (t(19) = -5.23, P < 0.001); Time-2: (t(19) = -5.23, P < 0.001). Critically, we observed significant interaction effect between Condition and Time (F(1,19) = 4.97, P = 0.038). Follow-up tests revealed significantly larger longitudinal improvement in accuracy in solving Addition (t(19) =

2.50, $P = 0.022$) relative to Control ($t(19) = 0.68$, $P = 0.50$) problems (**Fig. S1b**). We then conducted another 2-by-2 ANOVA for RT data revealed significant main effects of Condition ($F(1, 19) = 79.74$, $P < 0.001$) and Time ($F(1, 19) = 28.43$, $P < 0.001$). Follow-up tests revealed that children responded generally slower in solving Addition than Control problems (Time-1: ($t(19) = -9.29$, $P < 0.001$); Time-2: ($t(19) = -6.87$, $P < 0.001$), and they became faster at Time-2 than Time-1 for solving problems (Addition and Control: $t(19) > 4.63$, $P < 0.001$). Again, we observed significant interaction effect between Condition and Time ($F(1,19) = 5.67$, $P = 0.028$). Follow-up tests revealed significantly larger longitudinal improvement in RTs when solving Addition ($t(19) = 5.63$, $P < 0.001$) relative to Control ($t(19) = 4.34$, $P < 0.001$) problems (**Fig. S1d**).

Cross-sectional developmental changes in block and event-related fMRI task performance from childhood through adolescence into adulthood

We then examined cross-sectional changes in accuracy and RTs during solving addition problems across children at Time-2, adolescents and adults in the block and event-related fMRI tasks. We conducted a 2-by-3 ANOVA with Condition as within-subject factor and Group as between-subject factor for accuracy data from the block fMRI task. This revealed significant main effects of Condition ($F(1, 65) = 8.14$, $P < 0.001$) and Group ($F(2, 65) = 9.15$, $P < 0.001$). Post-hoc Scheffe's tests for the Group effect revealed significant cross-sectional improvement in accuracy from childhood through adolescence into adulthood, with higher accuracy in Adolescents ($P = 0.001$) and Adults ($P = 0.006$) than Children but no difference between Adolescents and Adults ($P = 0.90$). Critically, we observed significant interaction between Group and Condition ($F(2,65) = 6.26$, $P = 0.003$). Follow-up Scheffe's tests in two one-way ANOVAs separately for Addition and Control conditions revealed relatively larger cross-sectional improvement in solving Addition (Children < Adolescence: $P < 0.001$; Children < Adults: $P <$

0.001) relative to Control (Children < Adolescence: $P = 0.013$; Children < Adults: $P = 0.093$) problems from childhood through adolescence into adulthood (**Fig. S1a**).

Similarly, we conducted a 2-by-3 ANOVA for RT data from the block fMRI task. This analysis revealed significant main effects of Group ($F(2, 65) = 53.14$, $P < 0.001$) and Condition ($F(1, 65) = 56.72$, $P < 0.001$). Post-hoc tests using Scheffe's procedure revealed significant cross-sectional improvement in RTs from childhood through adolescence into adulthood, with faster RTs in Adolescents ($P < 0.001$) and Adults ($P < 0.001$) than Children but no difference between Adolescents and Adults ($P = 0.326$). Critically, we observed significant interaction between Group and Condition ($F(2,65) = 9.14$, $P < 0.001$). Post-hoc tests using Scheffe's procedure in two one-way ANOVAs separately for Addition and Control conditions revealed relatively larger cross-sectional improvement in RTs solving Addition (Children < Adolescence: $P < 0.001$; Children < Adults: $P < 0.001$) relative to Control problems (Children < Adolescence: $P = 0.001$; Children < Adults: $P = 0.001$) from childhood through adolescence into adulthood (**Fig. S1b**).

For the event-related fMRI task, we first conducted a 2-by-3 ANOVA for accuracy with Condition as within-subject factor and Group as between-subject factor. Consistent with accuracy in the block fMRI, we observed significant main effects of Group ($F(2, 57)$ 12, $P = 0.001$) and Condition ($F(1, 65) = 25.09$, $P < 0.001$). Post-hoc tests using Scheffe's procedure for the Group effect revealed significant cross-sectional improvement in accuracy from childhood through adolescence into adulthood, with higher accuracy in Adolescents ($P = 0.001$) and Adults ($P = 0.006$) than Children but no difference between Adolescents and Adults ($P = 0.90$). Critically, we observed significant interaction between Group and Condition ($F(2,65) = 6.26$, $P = 0.003$). Post-hoc tests using Scheffe's procedure in two one-way ANOVAs separately for Addition and Control conditions revealed relatively larger cross-sectional improvement in solving Addition (tests for Children < Adolescents: $P < 0.001$; Children < Adults: $P < 0.001$) relative to Control

(Scheffe's tests for Children < Adolescence: $P = 0.013$; Children < Adults: $P = 0.093$) problems from childhood through adolescence into adulthood (**Fig. S1c**).

Similarly, we conducted a 2-by-3 ANOVA for RT data from the event-related fMRI task. This analysis revealed significant main effects of Group ($F(2, 57) = 126.50$, $P < 0.001$) and Condition ($F(1, 57) = 178.21$, $P < 0.001$). Post-hoc tests using Scheffe's procedure revealed significant cross-sectional improvement in RTs from childhood through adolescence into adulthood, with faster RTs in Adolescents ($P < 0.001$) and Adults ($P < 0.001$) than Children, marginally faster RTs in Adolescents than Adults ($P = 0.092$). Critically, we observed very significant interaction between Group and Condition ($F(2,57) = 27.83$, $P < 0.001$). Post-hoc tests using Scheffe's procedure separately for Addition and Control conditions revealed relatively larger cross-sectional improvement in RTs solving Addition (Children < Adolescence: $P < 0.0001$; Children < Adults: P < 0.0001) relative to Control problems (Children < Adolescence: $P < 0.001$; Children < Adults: $P < 0.001$) from childhood through adolescence into adulthood (**Fig. S1d**).

Supplementary Tables: S1 to S9

Table S1: Participant demographics and neuropsychological assessments.

	Longitudinal fMRI		Cross-sectional fMRI		
	Children T1	Children T2	Adolescents	Adults	*P*
N (M/F)	28 (15/13)	28 (15/13)	20 (11/9)	20 (8/12)	0.40
Age (Range)	8.26 ± 0.53 (7 - 9)	9.45 ± 0.88 (9 - 11)	15.61 ± 1.40 (14 - 17)	20.50 ± 1.07 (19 - 22)	<0.001
Performance IQ	110.76 ± 12.57	-	112.35 ± 6.88	115.95 ± 10.28	0.32
Verbal IQ	112.13 ± 11.70	-	117.20 ± 10.12	122.25 ± 10.33	0.09
Full IQ	114.32 ± 8.93	-	116.80 ± 6.51	120.50 ± 8.85	0.10
Word reading	111.46 ± 12.41	108.04 ± 13.05	111.70 ± 4.62	111.58 ± 9.31	0.67
Math reasoning	110.57 ± 13.21	114.2 ± 12.71	113.7 ± 11.60	114.95 ± 8.45	0.38
Number operations	104.48 ± 13.10	109.37 ± 13.93	118.25 ± 9.45	116.47 ± 5.31	0.02

Mean (± standard deviation) of age, IQ, word reading, math reasoning and number operation measurements for all participants is shown. *P* values represent the significance of comparisons between children (Time-2), adolescent and adults. Notes: N, number of participants; T1, Time-1; T2, Time-2; - data not collected.

Table S2: Brain regions involved in addition problem solving in children.

Brain region	**R/L**	**BA**		***T* values**	**MNI (x, y, z)**
Addition > Control collapsing across Time-1 and Time-2					
Inferior frontal gyrus	L	47		4.24	-32 16 -2
	R			4.07	40 20 -8
Dorsolateral PFC	L	9		4.16	-38 20 26
	R			4.10	40 20 22
Insula	L	13		4.08	-38 14 10
	R			4.20	40 16 10
Dorsal ACC	L	6		3.76	-2 6 56
Hippocampus	**L**	**-**		**4.27**	**-20 -40 -4**
				2.93	**-20 -36 -2**
	R	**-**		**3.64**	**24 -38 0**
				2.74	**30 -20 -16**
Inferior parietal lobe	L	40		3.10	-46 -48 46
				3.06	-36 -60 44
Lingual gyrus	L	18		4.07	20 -100 0
Striatum	L	-		3.53	-18 -2 2
Midbrain	R	-		4.69	12 -30 -18
Cerebellum	L	-		3.22	-4 -52 -18

Only clusters, significant at a height threshold of $p < 0.01$ and an extent threshold of $p < 0.05$ corrected for multiple comparisons, are reported with local maxima in Montreal Neurological Institute (MNI) space. Clusters in the medial temporal lobe are in bold. Notes: BA, Brodmann's area; L, left hemisphere; R, right hemisphere; ACC, anterior cingulate cortex; PFC, prefrontal cortex; -, no proper data.

Table S3: Brain regions showing longitudinal changes in task-related activation during addition problem solving in children.

Brain regions	R/L	BA	*T* values	MNI (x, y, z)
		Children: Time-2 > Time-1		
Hippocampus	**L**	**-**	**3.64**	**-26 -24 -16**
			3.03	**-16 -20 -18**
	R		**3.90**	**28 -18 -18**
			3.72	**32 -14 -20**
Cerebellum	R	-	3.88	20 -68 -22
		Children: Time-1 < Time-2		
Dorsolateral PFC	L	8, 9	4.43	-44 18 52
			3.12	-32 30 50
	R		3.42	40 22 54
			3.40	48 14 36
Angular gyrus	R	39	4.00	44 -72 26
Superior parietal lobe	L	7	3.64	-22 -64 64

Notes are same as in Table S2.

Table S4: Brain regions showing longitudinal changes in task-related hippocampal functional connectivity and relations with individual gains in retrieval fluency in children.

Brain regions	R/L	BA	*T* values	MNI (x, y, z)
Functional connectivity: Children Time-2 > Time-1				
vmPFC	R	10	4.78	4 38 -2
			4.28	8 50 -2
Inferior frontal gyrus	L	47	3.44	-50 24 2
			3.19	-42 28 8
Anterior temporal lobe	L	38	3.92	-50 12 -10
			3.54	-54 16 -12
Positive correlation with individual gains in retrieval fluency in children from Time-1 to Time-2				
Dorsolateral PFC	R	46	5.09	46 32 40
			3.81	44 42 34
	L	46	3.87	-34 42 42
			3.84	-30 40 40
Inferior parietal sulcus	L	7	3.86	-30 -68 54
			3.17	-30 -76 50

Notes are same as in Table S2. PFC, dorsolateral prefrontal cortex; vmPFC, ventromedial prefrontal cortex.

Table S5: Longitudinal changes in task-related hippocampal functional connectivity with fronto-parietal regions *predict* individual gains in retrieval fluency in children.

Brain regions	*Correlation*	*Prediction*	
	r	$r_{(predicted, observed)}$	***p***
Left DLPFC	0.63	0.53	<0.001
Right DLPFC	0.80	0.71	<0.001
Left IPS	0.59	0.51	=0.001

Notes: DLPFC, dorsolateral prefrontal cortex; IPS, inferior parietal sulcus. “Correlation” refers to results from a conventional regression analysis of shared covariance between two variables. “Prediction” refers to the results from a machine learning algorithm with balanced 4-fold cross-validation combined with linear regression (see online Methods).

Table S6: Brain regions involved in addition problem solving in children, adolescents and adults.

Brain regions	R/L	BA	*T* values	MNI (x, y, z)
Children: Addition > Control				
(see Table S2 above)				
Adolescents: Addition > Control				
Inferior frontal gyrus	R	44	3.73	52 10 16
			3.65	50 20 14
Posterior inferior frontal gyrus	L	44	3.15	-38 4 26
Dorsal ACC	R	32	2.99	10 12 42
Insula	R	48	2.84	46 10 2
			2.87	36 20 -8
Supplementary motor area	L	6	2.80	-2 18 48
Inferior parietal lobe	L	7	2.80	-24 -66 48
Adults: Addition > Control				
Inferior frontal gyrus	R	44	3.00	42 12 24
	L		4.29	-46 6 28
Posterior inferior frontal gyrus	L	44	3.60	-42 8 36
Dorsal ACC	R	32	3.33	8 16 44
Supplementary motor area	R	6	3.00	2 18 48
Inferior parietal lobe	L	40, 7	3.58	-32 -54 36
			2.80	-24 -72 34

Notes are same as Table S2

Table S7: Brain regions showing developmental changes in task-related activation in children, adolescents and adults.

Brain regions	**R/L**	**BA**		***F* values**	**MNI (x, y, z)**
Omnibus *F* contrast: children at Time-2 vs. adolescents vs. adults					
Hippocampus	**R**	**-**		**8.46**	**32 -16 -18**
Cerebellum	R	-		8.27	14 -30 18

Notes are same as in Table S2

Table S8: Brain regions showing developmental changes in inter-problem multivoxel pattern stability from childhood through adolescence into adulthood.

Brain regions	R/L	BA	*F* values	MNI (x, y, z)
Omnibus *F* contrast: children at Time-2, adolescents versus adults				
Inferior frontal sulcus	L	44	11.21	-40 16 34
Inferior frontal gyrus		48	12.84	-54 20 14
Ventral-temporal cortex	L	37	14.74	-42 -38 -16
(fusiform)	R	19	15.69	40 -74 -14
Middle temporal gyrus	R	21	18.35	62 -32 -8
Postcentral gyrus	L	3	13.27	-40 -18 44
			8.96	-46 -22 36
Hippocampus	**L**	**-**	**12.68**	**-26 -14 -20**
			11.64	**-22 -32 -2**
	R	**-**	**9.60**	**20 -16 -16**
			9.22	**32 -6 -18**
Insula	L	48	10.29	-38 -8 -6
Calcarine sulcus	R	18	12.38	16 -66 18
Thalamus	L	-	12.46	-20 -30 0
	R	-	9.11	10 -22 -2
Midbrain	L	-	9.44	-8 -32 -10

Only clusters, significant at a height threshold of $P < 0.001$ and an extent threshold of $P < 0.05$ corrected for multiple comparisons, are reported with local maxima in MNI standard space. Other notes are same as Table S3.

Table S9: Original and matched number of correct problems included in matched multivoxel pattern stability analysis.

	Original (means ± standard deviations)	Matched in means and standard deviations
Children T1	19.35 ± 3.35	19.35 ± 3.35
Children T2	22.05 ± 2.35	19.35 ± 3.36
Adolescents	24.65 ± 1.42	19.45 ± 3.63
Adults	25.50 ± 0.69	19.50 ± 3.44

References

1. Friston, K.J., *et al.* Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* **6**, 218-229 (1997).

2. Sanchez, C.E., Richards, J.E. & Almli, C.R. Age-specific MRI templates for pediatric neuroimaging. *Developmental neuropsychology* **37**, 379-399 (2012).