Behavioral/Systems/Cognitive

A Three-Year Longitudinal Functional Magnetic Resonance Imaging Study of Performance Monitoring and Test-Retest Reliability from Childhood to Early Adulthood

P. Cédric M. P. Koolschijn,[1,2] Margot A. Schel,[1,2] Mark de Rooij,[1] Serge A. R. B. Rombouts,[1,2,3] and Eveline A. Crone[1,2,4]
[1]Institute of Psychology, Brain and Development Laboratory, Leiden University, 2333 AK Leiden, The Netherlands, [2]Leiden Institute for Brain and Cognition, 2300 RC Leiden, The Netherlands, [3]Department of Radiology, Leiden University Medical Center, 2300 RC Leiden, The Netherlands, and [4]Department of Developmental Psychology, University of Amsterdam, 1018 WB Amsterdam, The Netherlands

Previous cross-sectional functional magnetic resonance imaging studies have shown that performance monitoring functions continue to develop well into adolescence, associated with increased activation in brain regions important for cognitive control (prefrontal cortex, anterior cingulate cortex, and parietal cortex). To date, however, the development of performance monitoring has not yet been studied longitudinally, which leaves open the question whether changes can be detected within individuals over time. In the present study, human boys and girls, between ages 8 and 27 years, performed a child-friendly rule-switch task in the scanner on two occasions ~3.5 years apart. Change versus stability was examined using two methods: (1) repeated-measures analyses and (2) test-retest reliabilities of blood oxygenation level-dependent responses. Results showed that with increasing age, participants performed better on the task. The changes in neural activation associated with the processing of performance feedback were, however, more reliably correlated with changes in performance than with age. Test-retest reliability was at least fair to good for adults and adolescents, but poor to fair for the youngest age group. Substantially more variability was observed in the pattern and magnitude of children compared with adults, which may be interpreted as proxy for developmental change. Together, the results show that (1) change within individuals is variable, and more so for children than for adolescents and adults, and (2) performance is a better predictor for change in neural activation over time. These findings set the stage for studying developmental change in the perspective of multiple predictors, rather than solely by divisions based on age groups.

Introduction

Performance monitoring is one of the main components of successful learning. It involves the monitoring for and detection of errors and change signals, which then allows for adjustment of ongoing behavior to optimize subsequent performance (Holroyd and Coles, 2002). Previous studies indicate that performance monitoring improves steadily throughout development, but does not reach adult levels until late childhood or early adolescence (Bunge et al., 2002; Crone et al., 2008; Luna, 2009).

Evidence from functional magnetic resonance imaging (fMRI) studies shows that children, adolescents and adults recruit a similar network of brain areas during performance monitoring, including the lateral prefrontal cortex (LPFC), anterior cingulate cortex (ACC)/presupplementary motor area (preSMA), and parietal cortices. However, the pattern of activation differs between children and adults, such that there is an increase in activation in LPFC and parietal cortex following feedback indicating performance adjustment (Crone et al., 2008; van Duijvenvoorde et al., 2008; van den Bos et al., 2009; Tau and Peterson, 2010). These findings are consistent with results of other developmental neuroimaging studies, showing that children and adolescents have immature activation in the frontoparietal network when performing tasks requiring cognitive control (Bunge et al., 2002; Rubia et al., 2006; Velanova et al., 2008).

These previous studies have provided the building blocks for understanding the neural substrates involved in the development of cognitive control and performance monitoring. However, these studies were all cross-sectional, therefore only providing a proxy of development. To truly assess developmental trajectories, longitudinal studies are of paramount importance. Compared with cross-sectional research, longitudinal research has several advantages and provides additional information. First, longitudinal studies can overcome problems with differential sampling across age groups, masking of within-individual change by variability across individuals, and difficulty in identifying complex developmental trajectories (Kraemer et al., 2000). Second, longitudinal methods test for individual patterns of change rather than group differences. Third, longitudinal research has more power to detect small developmental differences in behavior and in task-related brain activation (Durston et al., 2006). Fourth, the test of

Received Dec. 9, 2010; revised Jan. 17, 2011; accepted Jan. 30, 2011.

This research was supported by a VIDI grant (no. 91786368) from the Netherlands Organization for Scientific Research (NWO) to E.A.C. We thank Dr. Silvia Bunge for reading and commenting on the manuscript. We are thankful to Rosa Meuwese and Sandy Overgaauw for their help with data acquisition.

The authors declare no competing financial interests.

Correspondence should be addressed to Dr. P. Cédric M. P. Koolschijn, Institute of Psychology; Brain and Development Laboratory, Leiden University, 3B33.A, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands. E-mail: koolschijnpcmp@fsw.leidenuniv.nl.

DOI:10.1523/JNEUROSCI.6415-10.2011

Table 1. Demographics at both time points

			Mean age in yrs [range]		Mean scan interval	Raven raw	Estimated
	n	Sex	TP1	TP2	(years)	scores (SD)	IQ (SD)
Adults	10	6F	20.67 [18–24]	24.05 [21–27]	3.38	53.0 (6.32)	119 (13.46)
Adolescents	12	6F	15.01 [14–15]	18.97 [17–19]	3.95	56.3 (2.83)	126 (7.15)
Children	10	4F	10.38 [8–11]	14.03 [11–15]	3.64	51.9 (4.01)	122 (7.04)

within-subjects change may contribute to the question how differences in brain activity are associated with age vis-à-vis performance changes over time.

The aim of this study was to examine within-subject changes in brain activity when performing a feedback-based rule-switching task using a longitudinal design in participants aged 8–27. The advantage of this approach is that it allows for the assessment of change across a wide age range using a relatively limited time scale (3.5 years). We used two methods to test change versus stability over time. First, changes were studied using repeated-measures analyses. Second, stability was examined using an accurate assessment of test-retest reliability of functional activity. Previous studies have demonstrated that fMRI signals provide relative reliable measures over sessions in adults (Bennett and Miller, 2010), but the reliability of fMRI signals has not yet been reported in children, or in studies with a time interval longer than one year. Thus, we tested both change (repeated measures) and stability (test-retest reliability of fMRI activation levels) over time in the same rule-switching task for all age groups. Our prediction was that a larger change in brain activation over time would be observed for the children relative to adolescents and adults (Ferrer et al., 2009). Consequently, we predicted that stability would be lower for children than for adolescents and adults.

Materials and Methods

Participants. A 3.5 year longitudinal functional magnetic resonance imaging study was carried out, including healthy adults, adolescents and children (see Table 1 for demographic information). In the full baseline sample of Crone and colleagues (Crone et al., 2008), 20 adults (12 females) age 18–25, 20 adolescents (9 females) age 14–15, and 17 children (8 females) age 8–11 were included in the study. Thirty-two participants; 10 adults (6 females), 12 adolescents (6 females) and 10 children (4 females) completed the longitudinal study and were rescanned after an interval of $\sim$3.5 years. The scan interval for adolescents was significantly longer compared with adults ($p = 0.01$), and there was no difference in scan interval length between children and adults ($p = 0.22$) or between children and adolescents ($p = 0.36$). The full baseline data have been published previously (Crone et al., 2008; Zanolie et al., 2008). Standard intelligence scores were obtained from each participant using the Raven's Progressive Matrices test (Raven et al., 1998). All estimated IQ scores were within the normal range and there were no significant differences between age groups ($F_{(2,29)} = 1.61$; $p = 0.22$) (Table 1).

The participants, all of whom received course credit or a fixed payment, were healthy right-handed volunteers with no history of neurological or psychiatric problems. Informed consent was obtained and the study was approved by the Internal Review Board at the Leiden University Medical Center.

Procedure and experimental design. All participants were tested individually and were trained to lie still in a mock scanner, which simulated the environment and sounds of an actual MRI scanner. Details of the child-friendly rule-switch task have been published previously (Crone et al., 2004, 2008; Zanolie et al., 2008). In short, participants were asked to respond to a stimulus that could appear in one of four horizontally presented locations on the screen by pressing appropriate buttons (Fig. 1). Before scanning, participants were trained to perform three spatial stimulus–response rules. Each response was followed by a positive or a negative feedback signal. Feedback stimuli were displayed as a "plus" (representing positive feedback) or "minus" (representing negative feedback) sign. Next, the participants were instructed that in the real experiment they had to find the correct rule themselves, which could be any of the three spatial rules they had just learned. They were instructed (1) to find the rule by using positive and negative feedback, (2) to apply that rule which would yield positive feedback, and (3) that the rules could change unexpectedly. Therefore, they had to apply the correct rule until a rule switch occurred, which was signaled by a negative feedback sign.

The four possible answers were mapped to the four buttons which were mapped to the index and middle fingers from the left and right hand (Fig. 1, top). Following rule 1, stimuli that appeared in one of the four locations designated a response with the finger compatible to the location. Thus, spatially compatible responses were required in response to the location of the stimulus. Following rule 2, stimuli that appeared in any of the four locations designated a response with the opposite finger of the same hand. Following rule 3, stimuli that appeared in any of the four locations designated a response with the finger that was assigned to the location two positions from the stimulus location (Fig. 1, bottom). Participants were told that rules could switch from time to time without a warning and they were instructed to use the trial-to-trial feedback to infer the correct response rule. Rules changed in a pseudorandomized order when participants had correctly applied the previous response rule for two to four consecutive trials. On each trial, a 2.5 s stimulus display was presented that required a button-press. If a response was not made within 2.5 s, a warning was presented indicating that the response was too slow and that faster responses were required on the next trial. These trials were not included in the analysis and consisted of $<1\%$ of the trials. When the participant responded within the 2500 ms timeframe, the feedback display consisted of two similar houses with four doors and a fixation mark, with an additional plus or minus sign placed in the door selected by the participant during the response time. After the feedback an intertrial interval jitter varying from 2000 to 8000 ms was presented in 25% of trials.

Feedback scoring. The five feedback types were determined *post hoc* for each individual separately. Their definitions were as follows: (1) First warning negative feedback was the first negative feedback that followed a successfully completed sequence of rule applications. This negative feedback was given unannounced once the rule had been correctly applied on two, three, or four consecutive trials (randomly determined for each rule separately); it indicated that the previously applied response rule was no longer correct, and thus indicated a rule switch. (2) Efficient negative feedback indicated that the rule chosen when searching for the appropriate rule was incorrect. After receiving efficient negative feedback participants had to apply the correct rule in the next trial. (3) Other error negative feedback trials consisted of those trials in which the participant failed to apply the correct response when the response rule had not changed. It also included those trials in which the participant perseverated in using the previously correct rule after a rule change. (4) First positive feedback indicated that the correct rule had been found following a rule switch, and (5) positive feedback indicated correct rule use.

Data acquisition. Trials were presented in three scans of 8.2 min each. During scanning, 300 trials were presented and the rules were switched in pseudorandomized order. The order of trials within each scan was determined by using an optimal sequencing program designed to maximize efficiency of recovery of the blood oxygenation level-dependent (BOLD) response (Dale, 1999). Scanning was performed with a standard whole-head coil on a Philips 3.0 Tesla scanner at the Leiden University Medical Center. Functional data were acquired using T2*-weighted echoplanar imaging (EPI) during three functional runs of 232 volumes each, of which the first 2 volumes were discarded to allow for equilibration of T1 saturation effects [repetition time = 2.211 s (2.2 s at follow-up), echo time = 30 ms, ascending interleaved acquisition, 38 slices of 2.75 mm, field of view 220 mm, 80 × 80 matrix, in-plane resolution 2.75 mm]. High resolution T1-weighted anatomical images were also collected after the functional runs. Head motion was restricted using a pillow and foam inserts that surrounded the head. Visual stimuli were projected onto a screen that was viewed through a mirror.

fMRI data analysis. All data (i.e., those who participated at the second measurement, and those who dropped out) (Crone et al., 2008) were reanalyzed using SPM5 (Wellcome Department of Cognitive Neurology, London, UK). Images were corrected for differences in timing of slice acquisition, followed by rigid body motion correction. Structural and functional volumes were spatially normalized to T1 and EPI templates, respectively. Translational movement parameters never exceeded 1 voxel (<3 mm) in any direction for any participant or scan. There were no significant differences in movement parameters between age groups [time point 1 (TP1): $F_{(2,29)} = 2.70$; $p = 0.10$; TP2: $F_{(2,29)} = 2.32$; $p = 0.12$]. The normalization algorithm used a 12-parameter affine transform together with a nonlinear transformation involving cosine basis functions and resampled the volumes to 3 mm cubic voxels. Templates were based on the MNI305 stereotaxic space (Cocosco et al., 1997), an approximation of Talairach space (Talairach and Tournoux, 1988). Functional volumes were spatially smoothed with an 8 mm full-width half-maximal isotropic Gaussian kernel. Statistical analyses were performed on individual subjects data using the general linear model in SPM5. The fMRI time series data were modeled by a series of events convolved with a canonical hemodynamic response function. The feedback stimulus of each trial was modeled as an event of interest. The trial functions were used as covariates in a general linear model, along with a basic set of cosine functions that high-pass filtered the data, and a covariate for session effects. The least-squares parameter estimates of height of the best-fitting canonical hemodynamic response function for each condition were used in pairwise contrasts. The resulting contrast images, computed on a subject-by-subject basis, were submitted to group analyses. Task-related responses were considered significant if they consisted of at least 10 contiguous voxels that exceeded a stringent threshold $p < 0.001$ [false discovery rate (FDR) corrected] (Genovese et al., 2002). Analyses were also performed with less stringent thresholds, $p < 0.01$ and $p < 0.05$, FDR and cluster corrected, but there were no differences compared with our stringent threshold.

In the fMRI analyses, we focused on the contrast first warning > positive feedback. This contrast provides the cleanest form of the processing of a change cue indicating performance adjustment relative to a low level task application baseline (positive feedback). Positive feedback indicated only those trials where the correct rule was applied, thereby excluding the first positive feedback-type due to possible novelty effects after finding the correct rule. Whole-brain, voxelwise between-group repeated measures analyses were performed for activation patterns associated with first warning versus positive feedback processing with and without adding age or performance at TP1 as a covariate.

Region of interest (ROI) analyses were performed to further characterize rule sensitivity of predicted brain regions based on the first measurements. ROI analyses were performed with the MarsBaR toolbox in SPM5 (Brett et

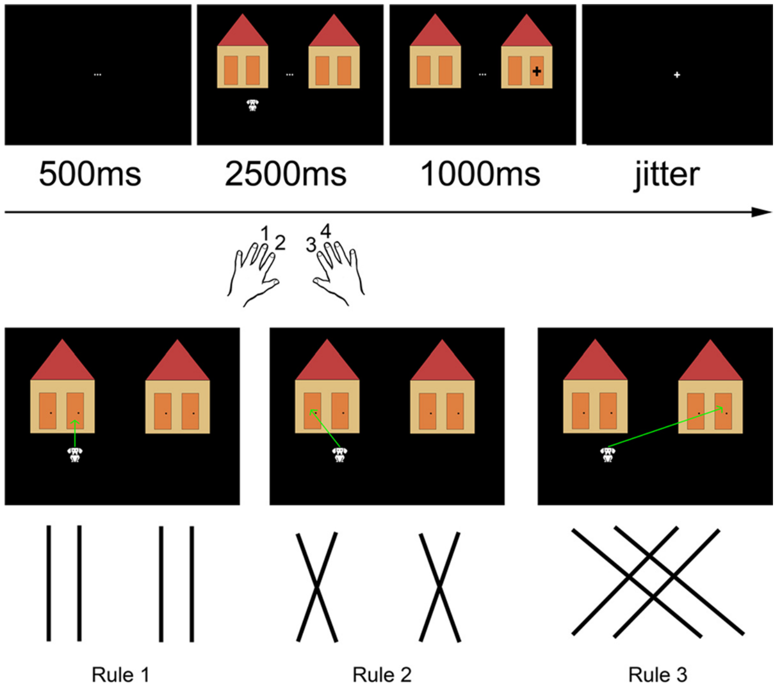

Figure 1. Display of task sequence (top) and rule types (bottom). Subjects were told to infer one of the spatial mapping rules that were trained before scanning. The task was changed into a prosocial game by explaining to the subjects that they should help the dog find its way back home. The dog could appear in one of four locations, and the subjects were instructed to open one of the four doors (locations) by pressing the corresponding response key. Their selection was followed by a visually presented feedback sign (+ or −). The rules changed unannounced following two, three, or four consecutive correct sorts. The spatial mapping rules are displayed in the bottom part the figure (see Procedure and experimental design for explanations).

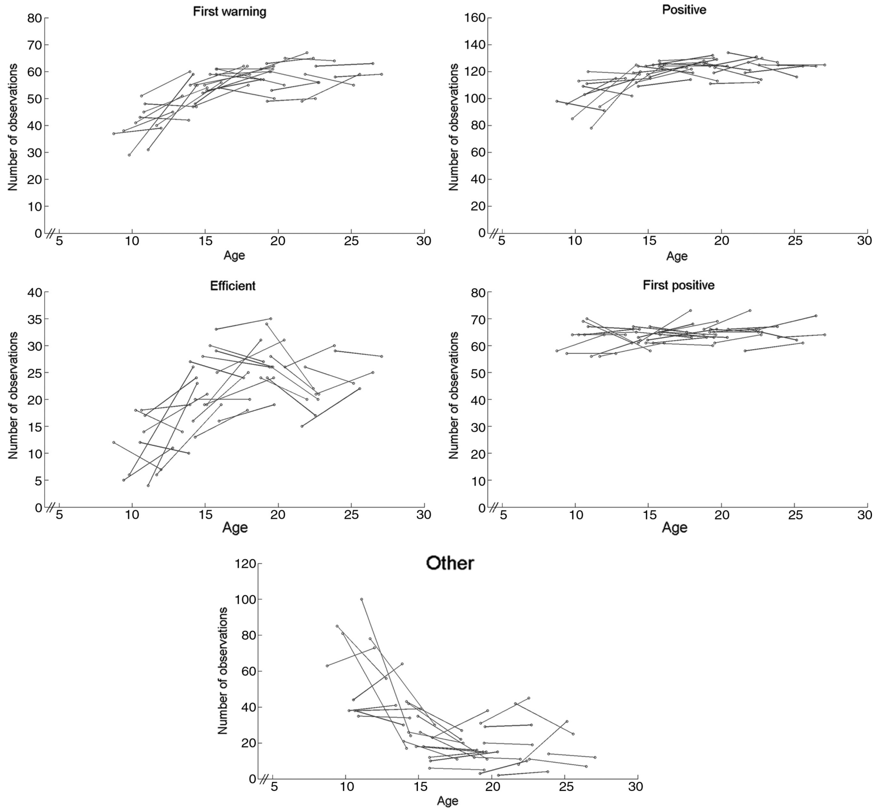

Figure 2. Number of feedback observations as a function of age for all feedback types. Each line represents an individual participant at time point 1 and time point 2.

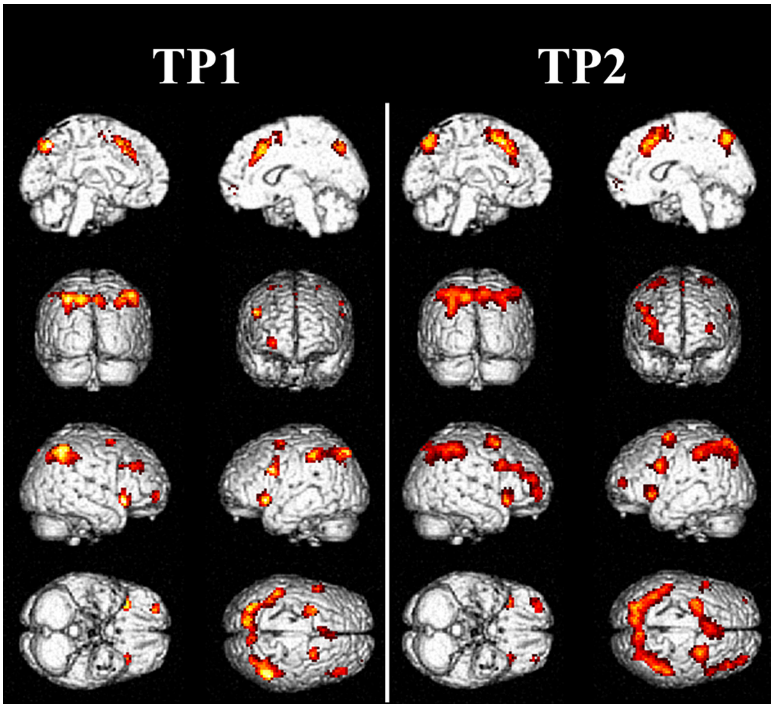

Figure 3. Cross-sectional feedback-locked whole-brain contrasts showing effects of first warning > positive feedback (FDR corrected, $p < 0.001$; 10 contiguous voxels). Group-activation patterns were similar across the two sessions except for slightly stronger overall activation on TP2 compared with TP1.

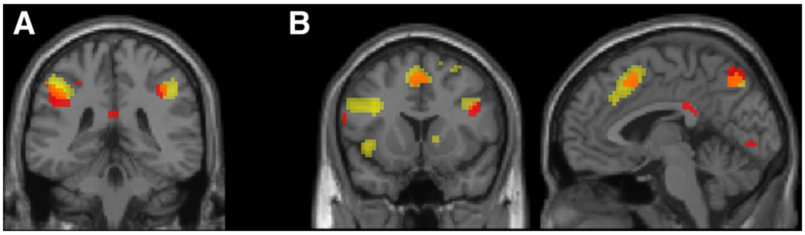

Figure 4. Feedback-locked whole-brain contrasts showing effects of first warning > positive feedback (yellow) and overlap with task performance (orange). ***A***, Overlap in bilateral parietal cortices ($y = -37$). ***B***, Overlap in LPFC and preSMA/ACC ($y = 13$). Both activation patterns are thresholded at $p < 0.001$, FDR correction, with at least 10 contiguous voxels.

al., 2002) (http://marsbar.sourceforge.net/). ROIs that spanned several functional brain regions were subdivided by sequentially masking the functional ROI with each of several anatomical MarsBaR ROIs. The contrast used to generate functional ROIs was based on a t test for first warning versus positive feedback stimuli based on the first measurement for only those participants who participated at the second measurement. For all ROI analyses, effects were considered significant at α of 0.001 [FDR and cluster corrected (at least 10 contiguous voxels)]. Results were similar with less stringent thresholds, therefore we only report findings with α of 0.001 Although there was a significant difference in scan interval between the three age groups (i.e., the scan interval was shorter for adults compared with adolescents), in none of the performed analyses there was a significant effect of scan interval. Therefore, all analyses are reported without scan interval as a covariate. There were also no significant differences in neural activity between those participants which dropped out after the first measurement and the participants that participated in the second scanning session.

Reliability measurements. Reliability of brain activation was analyzed by calculating intraclass correlation coefficients (ICCs). We calculated measures of intravoxel reliability on individual contrast values for each ROI by using the ICC toolbox provided by Caceres et al. (2009). The same ROIs were used as for the functional analyses. By analyzing only ROIs based on the first measurement we could test whether the level of group activation of the first session could predict the consistency in participant activations. Although no consensus has been achieved regarding reliability criteria for fMRI studies, previous studies have proposed different criteria. We followed the guidelines proposed by Cicchetti for quantifying reliability: poor (<0.4), fair (0.41–0.59), good (0.60–0.74) or excellent (>0.75) (Cicchetti and Sparrow, 1981; Cicchetti, 2001). These proposed criteria parallel suggested acceptance levels of the neuroimaging community of critical ICC values of 0.4 (Eaton et al., 2008) or 0.5 (Aron et al., 2006).

One problem with this method is that ICC measures can be biased due to between-subject variability. Although, the reported median ICC is thought to be more robust for between-subjects variance, another form of assessing reliability is to solely examine the within-subject SD (Zandbelt et al., 2008). The reliability metric reported is the SD of the change score ($\sigma_{\text{w-corrected}}$).

Table 2. Association between performance change over time and change in neural activation associated with first warning relative to positive feedback processing

Area[a]	Coordinates center of mass			Partial r with ΔFW[b]	p value[c]
L superior/middle FG	−25	−1	56	**0.596**	**<0.001**
L insula	−34	20	−3	0.277	0.131
L inferior parietal cortex	−37	−48	45	**0.617**	**<0.001**
L superior parietal cortex	−22	−70	48	**0.598**	**<0.001**
L precuneus	−7	70	47	**0.571**	**0.001**
R precuneus	9	−70	46	**0.633**	**<0.001**
L inferior FG/operculum	−43	9	30	0.290	0.114
R superior FG	25	3	58	**0.521**	**0.003**
R superior orbital gyrus	28	57	−3	**0.543**	**0.002**
R insula/inferior operculum	36	21	−4	0.319	0.081
R middle FG	45	31	32	**0.465**	**0.008**
R angular gyrus	38	−60	45	**0.662**	**<0.001**
R inferior parietal cortex	44	−49	47	**0.567**	**0.001**
R superior parietal cortex	39	−54	55	**0.562**	**0.001**
preSMA/MedFG/ACC	3	23	42	**0.596**	**<0.001**

L, Left; R, right; ΔFW, difference score of first warning feedback observations (TP2 − TP1); FG, frontal gyrus; Med, medial.

[a] Absolute ΔFW values for all ROIs were different from zero (all p values <0.001).

[b] Performance correlations were corrected for age at TP1.

[c] Significant correlations are printed in bold.

Results

All results reported below include only those participants who participated on both measurements.

Behavior

Performance differences between age groups were examined by comparing the number of feedback observations over the course of the experiment. An age group (8–11 years, 14–15 years, 18–24 years) by feedback type (first warning, efficient, other, first positive, and positive feedback) interaction showed that age groups differed in the number of feedback observations (age by feedback interaction, $F_{(8,116)} = 23.70$, $p < 0.001$). Comparisons for each feedback type separately showed that adolescents did not differ

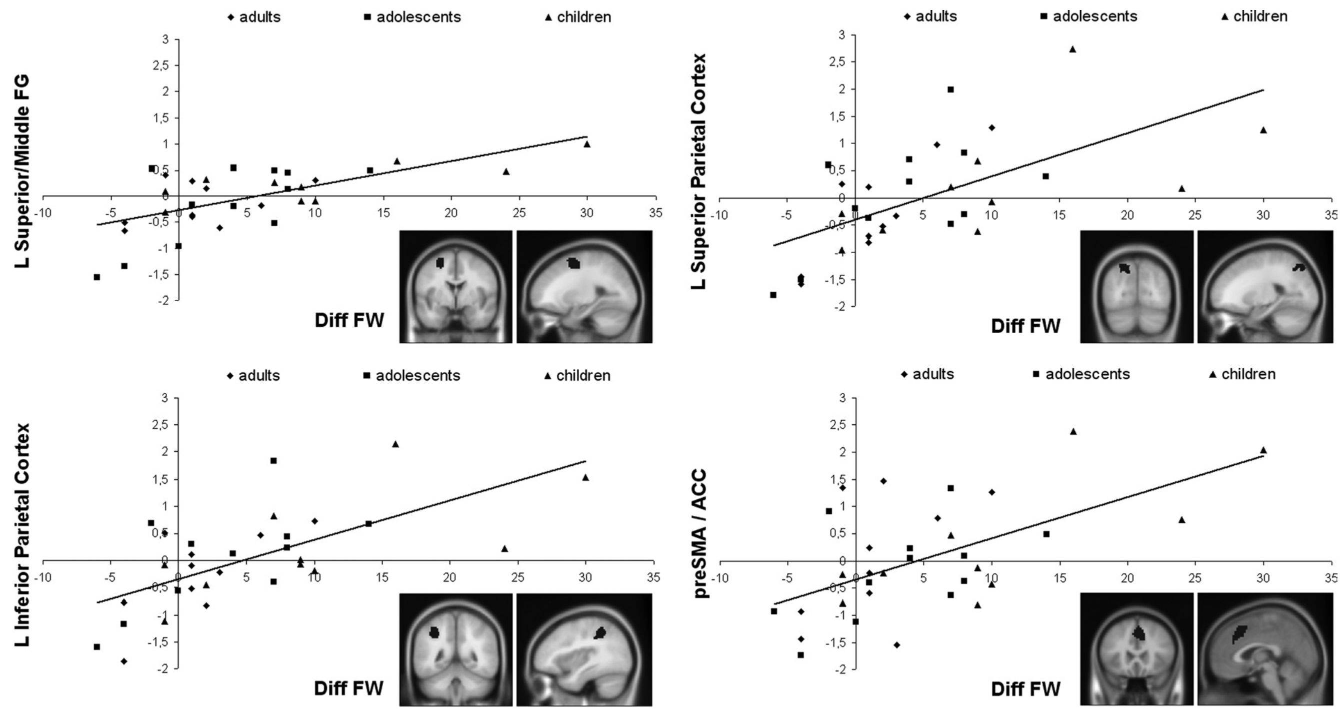

Figure 5. Correlations between performance change over time (Diff FW) and changes in neural activation were associated with first warning relative to positive feedback processing over time. A selection of brain regions with the strongest association is shown clockwise, starting top left: left superior/middle frontal gyrus (−25, −1, 56), left superior parietal cortex (−22, −70, 48), preSMA/ACC (3, 23, 42), and left inferior parietal cortex (−37, −48, 45) are shown. L, Left; FG, frontal gyrus.

from adults in performance. In contrast, children had fewer first warning (indicating rule switches), efficient and positive feedback observations, but more other error feedback compared with adolescents and adults (all p values <0.001). An age group (8–11 years, 14–15 years, 18–24 years) by time (TP1, TP2) by feedback type (first warning, efficient, error, first positive, and positive feedback) interaction indicated that over time children improved more compared with adolescents and adults ($F_{(8,116)} = 3.10, p < 0.05$).

To examine the within-subject changes in performance, difference scores (TP2 − TP1) were calculated for all feedback types for each participant separately (Fig. 2). One-sample t tests showed that the absolute difference scores for each feedback type were different from zero (all p values <0.01), indicating that participants' performance changed from TP1 to TP2. Regression analysis showed that age at TP1 was a predictor of the difference in number of first waning ($\beta = -0.46$, $p < 0.01$, observed power = 0.88, other error ($\beta = 0.37, p < 0.05$, observed power = 0.7) and positive feedback observations ($\beta = -0.41$, $p < 0.05$, observed power = 0.81), indicating that the within-person improvement was larger for younger children than for older children and adults.

Whole-brain analyses

Whole-brain activation patterns were examined cross-sectionally at both time points using one-sample t tests (FDR corrected, $p <$ 0.001, 10 contiguous voxels). The results showed similar whole-brain activation patterns at TP1 and TP2 (Fig. 3), including activation in preSMA (middle) frontal regions and bilateral parietal cortices. These findings indicate that participants of all age groups recruited largely overlapping brain regions at baseline as well as at the follow-up session. An analysis with performance as a predictor at TP1 indicated that the same areas (LPFC, preSMA/ACC, and bilateral parietal cortex) were correlated with task performance at TP1, $p < 0.001$, FDR corrected with at least 10 contiguous voxels (Fig. 4).

Longitudinal analyses

First, we tested change in neural activation by performing a repeated-measures ANOVA for the contrast first warning > positive feedback, directly comparing TP1 and TP2, across all participants. This analysis did not show significant change over time. Second, we added age as a linear and log-linear covariate to the analyses, but again no significant changes were found. We reran the analyses with only the children and the adolescents (whole group and per subsample), since we expected the adults to show the least change and therefore they could bias the sample. Again, no significant changes in brain activity were found.

ROI analyses

ROI selection was based on the first warning > positive feedback contrast at baseline and included a wide range of areas, including preSMA/ACC, inferior and superior parietal cortices, bilateral superior/middle frontal gyrus, bilateral inferior frontal gyrus/inferior operculum BA44, and bilateral insula.

First, to examine within-subject changes in neural activation over time we calculated difference scores for all ROIs for each participant separately. One-sample t tests showed that the absolute difference scores for all ROIs were different from zero (all p values <0.001), indicating that participants' neural activation associated with first warning feedback processing changed from TP1 to TP2. However, the direction of activation change differed across participants. Regression analyses showed that age at TP1 was not a significant predictor of the within-subject change in feedback-related activation.

Second, we investigated the association between difference scores in performance (i.e., number of rule shifts) with differences scores of activation over time in the selected ROIs. Table 2

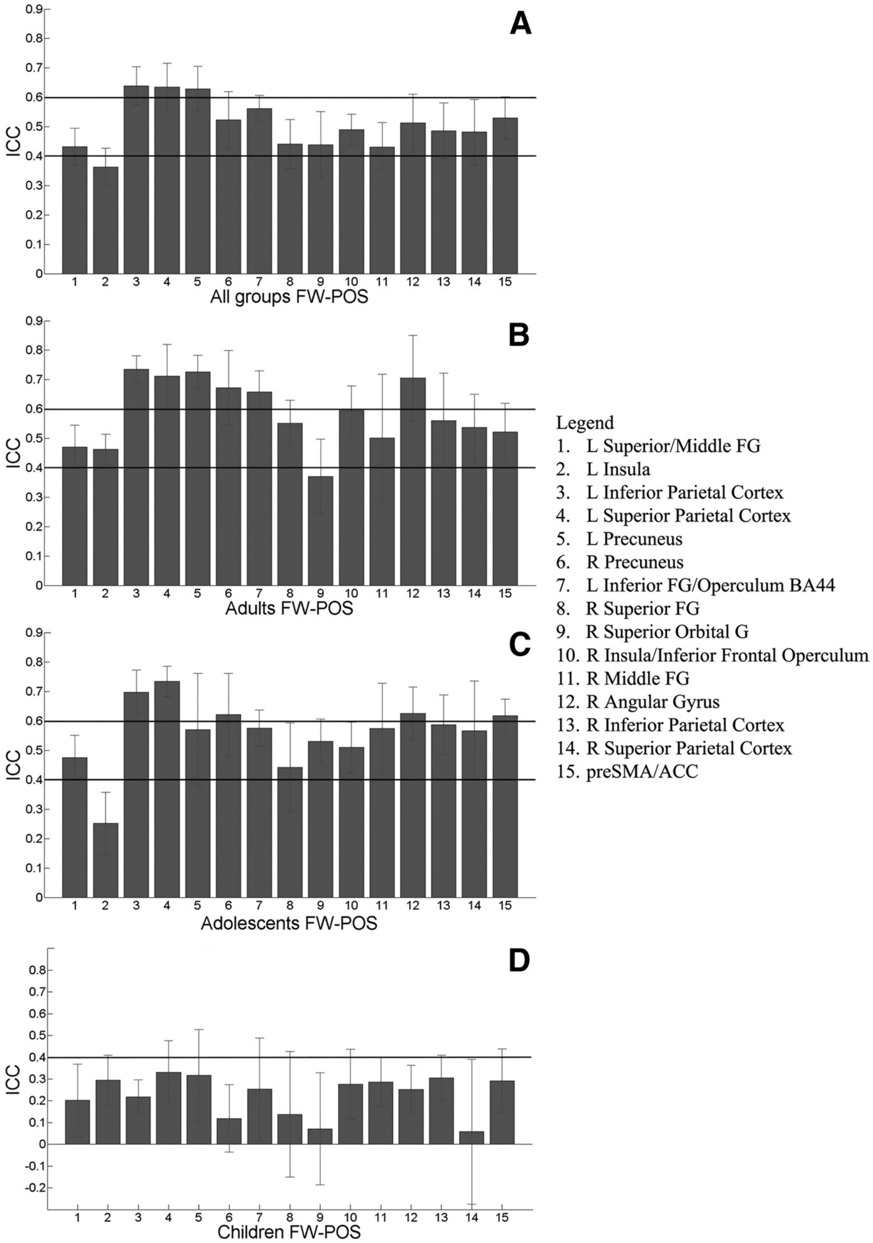

Figure 6. Intra-voxel reliability (ICC) measures based on ROIs at time point 1 (FDR corrected, $p < 0.001$, 10 contiguous voxels). ICCs were computed for each participant and population estimate was based on bootstrap methods. ***A*** displays ICC values with SE bands for the whole sample. In ***B–D***, the bars indicate ICC values for each age group: adults (***B***), adolescents (***C***), and children (***D***).

shows that change in the number of rule shifts was significantly positively associated with change in neural activation related to first warning negative feedback, even when age at TP1 was entered as a covariate to the analysis. These findings are also illustrated in Figure 5, and demonstrate that performance is a better predictor of neural changes over time than age. Notably, change in brain activity in bilateral middle frontal regions and left parietal cortex, brain areas previously associated with rule-switching, showed the highest correlations with performance change.

Testing for stability: Reliability measurements

The reliabilities of the first warning > positive feedback contrast for all ROIs were consistently in the fair to good range for the whole group (Fig. 6*A*; Table 3). Since we expected differences in reliability between the different age groups, intravoxel reliabilities were also calculated for each group separately (Fig. 6*B–D*; Table 3) (plots for each ROI per participant of the voxel values for TP1 against TP2 are available upon request). Adults and adolescents both showed at least fair reliabilities and good reliabilities for bilateral precuneus, left inferior and superior parietal cortices and right angular gyrus. In contrast, children demonstrated poor ICCs and higher SE bands in all regions compared with the older groups.

Since children performed less well on the task, it could be argued that lower reliability estimates in the ROIs are due to different (and lower) activation patterns. Therefore, we also analyzed ICCs for the contrast negative > positive feedback, in which the total observations of negative (first warning, efficient and other error feedback types) and positive (positive and first positive feedback types) feedback are similar across all age groups. For these analyses we redefined the ROIs specific for this contrast. Again, ROIs were based on TP1, FDR corrected, $p < 0.001$ with at least 10 contiguous voxels. There was considerable overlap between the ROIs from both contrasts (these ICC values are available upon request). Compared with the ICC values for first warning > positive feedback, values tended to be higher for all groups in the negative > positive feedback contrast. More specifically, the two youngest age groups showed higher reliability values up to one- to two-tenths for all ROIs averaged.

Finally, within-subject variability was also calculated in an attempt to protect against bias from between-subjects variance (Table 3). Similar to the ICC measurements, σ_w values are given for the whole sample and age groups separately. Since there were large differences in height of activation between the age groups, we corrected the individual σ_w values for mean activation per group across sessions and values are reported as $\sigma_{\text{w-corrected}}$. Although, $\sigma_{\text{w-corrected}}$ values vary widely across brain regions, adults show overall the least within-subject variability compared with the adolescents and children, with children showing the largest within-subject variability. To create again equal number of trials for each age group, the $\sigma_{\text{w-corrected}}$ values were also calculated for the negative > positive feedback contrast. Similar as the ICC values, less within-subject variability was observed in the negative > positive feedback contrast compared with the first warning > positive feedback contrast (these $\sigma_{\text{w-corrected}}$ values are available upon request).

Discussion

The aim of this three year longitudinal study was to investigate developmental and within-subject changes in task perfor-

Table 3. Reliability measurements of ROIs for first warning > positive feedback contrast

	MNI coordinates			medICC (SE)				$\sigma_{\text{w-corrected}}$ (SE)			
Brain area	center of mass			All	Adults	Adolescents	Children	All	Adults	Adolescents	Children
L superior/middle FG	−25	−1	56	0.43 (0.06)	0.47 (0.07)	0.48 (0.08)	0.20 (0.17)	0.52 (0.06)	0.35 (0.04)	0.60 (0.08)	0.58 (0.07)
L insula	−34	20	−3	0.36 (0.06)	0.46 (0.05)	0.25 (0.11)	0.30 (0.11)	0.48 (0.06)	0.43 (0.05)	0.57 (0.07)	0.43 (0.05)
L inferior parietal cortex	−37	−48	45	0.64 (0.07)	0.73 (0.05)	0.70 (0.08)	0.22 (0.08)	0.56 (0.07)	0.33 (0.04)	0.42 (0.05)	0.96 (0.12)
L superior parietal cortex	−22	−70	48	0.64 (0.08)	0.71 (0.11)	0.73 (0.05)	0.33 (0.15)	0.57 (0.07)	0.35 (0.04)	0.43 (0.05)	0.96 (0.12)
L precuneus	−7	70	47	0.63 (0.08)	0.73 (0.06)	0.57 (0.19)	0.32 (0.21)	0.67 (0.08)	0.35 (0.04)	0.58 (0.07)	1.12 (0.14)
R precuneus	9	−70	46	0.52 (0.10)	0.67 (0.13)	0.62 (0.14)	0.12 (0.15)	0.93 (0.12)	0.34 (0.04)	0.45 (0.06)	2.08 (0.26)
L inferior FG/operculum	−43	9	30	0.56 (0.05)	0.66 (0.07)	0.58 (0.06)	0.25 (0.23)	0.51 (0.06)	0.37 (0.05)	0.56 (0.07)	0.58 (0.07)
R superior FG	25	3	58	0.44 (0.08)	0.55 (0.08)	0.44 (0.15)	0.14 (0.29)	0.56 (0.07)	0.39 (0.05)	0.41 (0.05)	0.91 (0.11)
R superior orbital gyrus	28	57	−3	0.44 (0.11)	0.37 (0.13)	0.53 (0.07)	0.07 (0.26)	0.69 (0.09)	0.55 (0.07)	0.82 (0.10)	0.68 (0.09)
R insula/inferior frontal operculum	36	21	−4	0.49 (0.05)	0.60 (0.08)	0.51 (0.09)	0.28 (0.16)	0.43 (0.05)	0.54 (0.07)	0.33 (0.04)	0.45 (0.06)
R middle FG	45	31	32	0.43 (0.08)	0.50 (0.22)	0.57 (0.15)	0.29 (0.11)	0.74 (0.09)	0.53 (0.07)	0.48 (0.06)	1.25 (0.16)
R angular gyrus	38	−60	45	0.51 (0.10)	0.71 (0.15)	0.63 (0.09)	0.25 (0.11)	0.69 (0.09)	0.45 (0.06)	0.49 (0.06)	1.17 (0.15)
R inferior parietal cortex	44	−49	47	0.49 (0.09)	0.56 (0.16)	0.59 (0.10)	0.31 (0.10)	0.69 (0.09)	0.42 (0.05)	0.39 (0.05)	1.33 (0.17)
R superior parietal cortex	39	−54	55	0.48 (0.11)	0.54 (0.11)	0.57 (0.17)	0.06 (0.33)	0.82 (0.10)	0.64 (0.08)	0.43 (0.05)	1.46 (0.18)
preSMA/ACC	3	23	42	0.53 (0.07)	0.52 (0.10)	0.62 (0.06)	0.29 (0.15)	0.60 (0.08)	0.61 (0.08)	0.38 (0.05)	0.86 (0.11)

medICC, Median intraclass correlation coefficient; L, left; R, right; FG, frontal gyrus.

mance and feedback-related neural activation patterns. We tested this question using two types of analyses; repeated-measures ANOVAs and test-retest reliability of fMRI activation levels over time using a rule-switching task. Two principal findings emerged from this study. First, while on the behavioral level participants performed more rule switches with increasing age; change in performance (i.e., number of rule-shifts) was a better predictor for change in activity over time than age. Second, test-retest reliability was at least fair to good for adults and adolescents, and poor for the youngest age group. Substantially more variability was observed in the pattern and magnitude of children compared with adults, which may be interpreted as proxy for developmental change.

Change in performance and feedback-related activity over time

Behaviorally, task performance improved as indexed by an increase in rule shifts with increasing age. This is in line with previous research showing that the ability to switch between rules continues to develop through childhood and adolescence (Crone et al., 2006; Somsen, 2007; Kalkut et al., 2009; Anokhin et al., 2010). Moreover, the increase in rule switching over time was greatest for the youngest participants, indicating that the ability to switch between rules shows a large developmental increase during childhood and adolescence, and continues to develop throughout adolescence. These findings are in accordance with those of Kalkut et al. (2009) indicating that set-shifting abilities develop from late childhood to early adulthood. Other behavioral longitudinal studies have reported change in early childhood and adolescence in a variety of cognitive domains (Thomas et al., 1999; Ferrer et al., 2007; Kail and Ferrer, 2007; De Brauwer and Fias, 2009; Luna, 2009).

On both time points participants recruited a network of brain regions, including lateral PFC, preSMA/ACC, inferior and superior parietal cortex, and bilateral insula when processing first warning negative feedback (Crone et al., 2008; van Duijvenvoorde et al., 2008; van den Bos et al., 2009; Tau and Peterson, 2010). Here, we tested the predictive value of age and performance on within-subject changes in feedback-related neural activation longitudinally. In contrast to the behavioral findings, the fMRI data indicated that age was not a good predictor of change in brain activity between the two sessions. However, we found significant correlations between performance and activation change over time in the feedback processing network, including bilateral middle frontal gyrus, left inferior and superior parietal cortices and preSMA/ACC. These findings suggest that performance is a better predictor of activation change over time than age. This is in line with previous research indicating that performance can explain age-related differences in neural activation (Bunge et al., 2002; Booth et al., 2004). The following question that needs to be answered is: Does increased brain activity lead to increased performance or vice versa? This question also raises the issue of neural efficiency: less is more, more is less or more is more. Although there are no suitable answers yet for both questions (Poldrack, 2010), our findings build upon recent developmental studies in which increased activation patterns were associated with increased performance (Klingberg et al., 2002; Olesen et al., 2003; Vannest et al., 2010).

The current findings further implicate that future developmental neuroimaging studies should take task performance into account when examining age-related differences in neural activation. To overcome these differences, one option would be to match age groups based on performance scores (Schlaggar et al., 2002). Another alternative would be to use a parametric manipulation of task difficulty. This kind of manipulation allows for *post hoc* comparisons between age groups at different levels of task difficulty, while controlling for task performance (Durston and Casey, 2006). The relationship between performance and neural activation is complicated and might differ between cognitive domains (Booth et al., 2004). The findings of this study show that age in years may not be the best predictor for developmental change, rather, future studies should examine the possibility of characterizing individuals according to the way they perform complex tasks and learn information (Schmittmann et al., 2006).

Test-retest reliability

Change versus stability was further tested using test-retest reliability in children, adolescents and adults. Two methods were used to assess reliability in this study. First, ICC values indicated that there were differences in reliability between the age groups. Adults and adolescents showed good reliabilities for the inferior and superior parietal cortices, bilateral precuneus and right angular gyrus. Fair to good reliabilities were found for the other areas. However, the youngest age group showed poor test-retest reliability for all ROIs. The differences in reliability could not be explained by differences in the number of observations (see also Genovese et al., 1997; Friedman and Glover, 2006).

Second, we used within-subject variation in fMRI signal changes across measurements to protect against between-subjects variance. Considerable within-subject variation in fMRI signal changes was found in the brain areas associated with the first warning > positive feedback contrast. There was a similar pattern compared with the ICC measurements, showing relatively low within-subject variation in adults (44% mean for all areas) compared with adolescents (49%) and children (99%). When within-subjects variance for the negative > positive feedback contrast was calculated, the variance reduced similar to the ICC findings.

Both measurements indicate that even after a 3.5 year scan interval reliability is fair to good for adults and adolescents, but this is accompanied with considerable within-subject variation. Although the youngest group showed low ICC values and large within-subject variation, this does not necessarily imply low reliability. In our opinion, these findings can also be interpreted in terms of ongoing maturational processes. Previous research has shown that the human brain continues to mature through early adulthood (Gogtay et al., 2004; Giedd, 2008) and this maturation is hypothesized to influence both performance and functional activation (Blakemore, 2008; Casey et al., 2008). Therefore, if there are ongoing functional and structural changes in the youngest age ranges, then it is reasonable to expect that reliability measurements will show weak results. This suggests that the traditional test-retest reliability analyses could serve as a proxy for development. Hence, recruitment of brain areas can be quantified in similar or different activational patterns within brain regions. Different usage will lead to low reliability measurements, possibly indicative for different strategy use and/or ongoing maturational processes. Additional support for this conclusion can be found from the individual intravoxel reliability data, in which large differences were shown between activity patterns in the ROIs between both time points especially for the younger age groups.

Although several studies have investigated the reliability and reproducibility of fMRI over time (for review, see Bennett and Miller, 2010), it has to be noted that the scan intervals that have been studied vary extensively and are either on the order of a few days, weeks or one year and focused mainly on young and middle aged (healthy) adults. Therefore, it is difficult to compare our reliability findings with previous studies.

Limitations and Future directions

Reliability can be affected by technical, physiological, and psychological factors. In part we controlled for this, by using exactly the same paradigm, scanner and scanning protocol (including mock-scanner training) on both measurements. In addition, it is unlikely that our study suffered from learning-effects, because al participants were trained to learn the three rules before scanning. Moreover, test-retest reliability of fMRI BOLD also depends on several psychophysiological effects such as changes in arousal, attention, fatigue, task acquaintance, and heart rate, and these might also contribute to an increased variability and therefore lower reproducibility. Despite these difficulties, fMRI results were satisfactorily reliable.

So far, the number of longitudinal functional neuroimaging studies pale compared with the vast amount of cross-sectional studies. Only two studies examined within-subject changes in a cognitive control task over time in 9-year-old children (Durston et al., 2006) and 15-year-old adolescents (Finn et al., 2010). The current study is the first to study changes in a much wider age range and with a larger sample. Together with the findings from this study, both studies showed that the longitudinal findings provide additional information and differences in activation patterns compared with the cross-sectional findings. Age and performance effects can be assessed in either a cross-sectional or longitudinal fashion. Cross-sectional imaging studies provide insights into age and/or performance differences in brain function; however, longitudinal studies are required to provide a true measure of the functional change over time.

In summary, the present 3.5 year longitudinal fMRI study in healthy children, adolescents and young adults provided important evidence of behavioral and brain activity-related change over time and test-retest reliability of fMRI in young age groups. The most notable finding was that performance on a feedback-based rule-switching task is a better predictor than age of changes over time in feedback-related brain activation. A next step in developmental neuroimaging work could be to characterize participants based on learning types, as these may be stronger predictors of neural change than age alone.

References

Anokhin AP, Golosheykin S, Grant J, Heath AC (2010) Developmental and genetic influences on prefrontal function in adolescents: a longitudinal twin study of WCST performance. Neurosci Lett 472:119–122.

Aron AR, Gluck MA, Poldrack RA (2006) Long-term test-retest reliability of functional MRI in a classification learning task. Neuroimage 29:1000–1006.

Bennett CM, Miller MB (2010) How reliable are the results from functional magnetic resonance imaging? Ann N Y Acad Sci 1191:133–155.

Blakemore SJ (2008) The social brain in adolescence. Nat Rev Neurosci 9:267–277.

Booth JR, Burman DD, Meyer JR, Trommer BL, Davenport ND, Parrish TB, Gitelman DR, Mesulam MM (2004) Brain-behavior correlation in children depends on the neurocognitive network. Hum Brain Mapp 23:99–108.

Brett M, Anton JL, Valabregue R, Poline JB (2002) Region of interest analysis using an SPM toolbox. Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2–6, 2002, Sendai, Japan.

Bunge SA, Dudukovic NM, Thomason ME, Vaidya CJ, Gabrieli JD (2002) Immature frontal lobe contributions to cognitive control in children: evidence from fMRI. Neuron 33:301–311.

Caceres A, Hall DL, Zelaya FO, Williams SC, Mehta MA (2009) Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage 45:758–768.

Casey BJ, Getz S, Galvan A (2008) The adolescent brain. Dev Rev 28:62–77.

Cicchetti DV (2001) The precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements. J Clin Exp Neuropsychol 23:695–700.

Cicchetti DV, Sparrow SA (1981) Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am J Ment Defic 86:127–137.

Cocosco CA, Kollokian V, Kwan RKS, Evans AC (1997) BrainWeb: online interface to a 3D MRI simulated brain database. Neuroimage 5:S425.

Crone EA, Ridderinkhof KR, Worm M, Somsen RJ, van der Molen MW (2004) Switching between spatial stimulus-response mappings: a developmental study of cognitive flexibility. Dev Sci 7:443–455.

Crone EA, Donohue SE, Honomichl R, Wendelken C, Bunge SA (2006) Brain regions mediating flexible rule use during development. J Neurosci 26:11239–11247.

Crone EA, Zanolie K, Van Leijenhorst L, Westenberg PM, Rombouts SA (2008) Neural mechanisms supporting flexible performance adjustment during development. Cogn Affect Behav Neurosci 8:165–177.

Dale AM (1999) Optimal experimental design for event-related fMRI. Hum Brain Mapp 8:109–114.

De Brauwer J, Fias W (2009) A longitudinal study of children's performance on simple multiplication and division problems. Dev Psychol 45:1480–1496.

Durston S, Casey BJ (2006) What have we learned about cognitive development from neuroimaging? Neuropsychologia 44:2149–2157.

Durston S, Davidson MC, Tottenham N, Galvan A, Spicer J, Fossella JA, Casey BJ (2006) A shift from diffuse to focal cortical activity with development. Dev Sci 9:1–8.

Eaton KP, Szaflarski JP, Altaye M, Ball AL, Kissela BM, Banks C, Holland SK (2008) Reliability of fMRI for studies of language in post-stroke aphasia subjects. Neuroimage 41:311–322.

Ferrer E, McArdle JJ, Shaywitz BA, Holahan JM, Marchione K, Shaywitz SE (2007) Longitudinal models of developmental dynamics between reading and cognition from childhood to adolescence. Dev Psychol 43:1460–1473.

Ferrer E, O'Hare ED, Bunge SA (2009) Fluid reasoning and the developing brain. Front Neurosci 3:46–51.

Finn AS, Sheridan MA, Kam CL, Hinshaw S, D'Esposito M (2010) Longitudinal evidence for functional specialization of the neural circuit supporting working memory in the human brain. J Neurosci 30:11062–11067.

Friedman L, Glover GH (2006) Reducing interscanner variability of activation in a multicenter fMRI study: controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. Neuroimage 33:471–481.

Genovese CR, Noll DC, Eddy WF (1997) Estimating test-retest reliability in functional MR imaging. I: Statistical methodology. Magn Reson Med 38:497–507.

Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15:870–878.

Giedd JN (2008) The teen brain: insights from neuroimaging. J Adolesc Health 42:335–343.

Gogtay N, Giedd JN, Lusk L, Hayashi KM, Greenstein D, Vaituzis AC, Nugent TF 3rd, Herman DH, Clasen LS, Toga AW, Rapoport JL, Thompson PM (2004) Dynamic mapping of human cortical development during childhood through early adulthood. Proc Natl Acad Sci U S A 101:8174–8179.

Holroyd CB, Coles MG (2002) The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol Rev 109:679–709.

Kail RV, Ferrer E (2007) Processing speed in childhood and adolescence: longitudinal models for examining developmental change. Child Dev 78:1760–1770.

Kalkut EL, Han SD, Lansing AE, Holdnack JA, Delis DC (2009) Development of set-shifting ability from late childhood through early adulthood. Arch Clin Neuropsychol 24:565–574.

Klingberg T, Forssberg H, Westerberg H (2002) Increased brain activity in frontal and parietal cortex underlies the development of visuospatial working memory capacity during childhood. J Cogn Neurosci 14:1–10.

Kraemer HC, Yesavage JA, Taylor JL, Kupfer D (2000) How can we learn about developmental processes from cross-sectional studies, or can we? Am J Psychiatry 157:163–171.

Luna B (2009) Developmental changes in cognitive control through adolescence. Adv Child Dev Behav 37:233–278.

Olesen PJ, Nagy Z, Westerberg H, Klingberg T (2003) Combined analysis of DTI and fMRI data reveals a joint maturation of white and grey matter in a fronto-parietal network. Brain Res Cogn Brain Res 18:48–57.

Poldrack RA (2010) Interpreting developmental changes in neuroimaging signals. Hum Brain Mapp 31:872–878.

Raven J, Raven JC, Court JH (1998) Manual for Raven's progressive matrices and vocabulary scales. Section 1: General overview. San Antonio, TX: Harcourt Assessment.

Rubia K, Smith AB, Woolley J, Nosarti C, Heyman I, Taylor E, Brammer M (2006) Progressive increase of frontostriatal brain activation from childhood to adulthood during event-related tasks of cognitive control. Hum Brain Mapp 27:973–993.

Schlaggar BL, Brown TT, Lugar HM, Visscher KM, Miezin FM, Petersen SE (2002) Functional neuroanatomical differences between adults and school-age children in the processing of single words. Science 296:1476–1479.

Schmittmann VD, Visser I, Raijmakers ME (2006) Multiple learning modes in the development of performance on a rule-based category-learning task. Neuropsychologia 44:2079–2091.

Somsen RJ (2007) The development of attention regulation in the Wisconsin Card Sorting Task. Dev Sci 10:664–680.

Talairach J, Tournoux P (1988) Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging. New York: Thieme.

Tau GZ, Peterson BS (2010) Normal development of brain circuits. Neuropsychopharmacology 35:147–168.

Thomas H, Lohaus A, Kessler T (1999) Stability and change in longitudinal water-level task performance. Dev Psychol 35:1024–1037.

van den Bos W, Güroğlu B, van den Bulk BG, Rombouts SA, Crone EA (2009) Better than expected or as bad as you thought? The neurocognitive development of probabilistic feedback processing. Front Hum Neurosci 3:52.

van Duijvenvoorde AC, Zanolie K, Rombouts SA, Raijmakers ME, Crone EA (2008) Evaluating the negative or valuing the positive? Neural mechanisms supporting feedback-based learning across development. J Neurosci 28:9495–9503.

Vannest J, Rasmussen J, Eaton KP, Patel K, Schmithorst V, Karunanayaka P, Plante E, Byars A, Holland S (2010) FMRI activation in language areas correlates with verb generation performance in children. Neuropediatrics 41:235–239.

Velanova K, Wheeler ME, Luna B (2008) Maturational changes in anterior cingulate and frontoparietal recruitment support the development of error processing and inhibitory control. Cereb Cortex 18:2505–2522.

Zandbelt BB, Gladwin TE, Raemaekers M, van Buuren M, Neggers SF, Kahn RS, Ramsey NF, Vink M (2008) Within-subject variation in BOLD-fMRI signal changes across repeated measurements: quantification and implications for sample size. Neuroimage 42:196–206.

Zanolie K, Van Leijenhorst L, Rombouts SA, Crone EA (2008) Separable neural mechanisms contribute to feedback processing in a rule-learning task. Neuropsychologia 46:117–126.