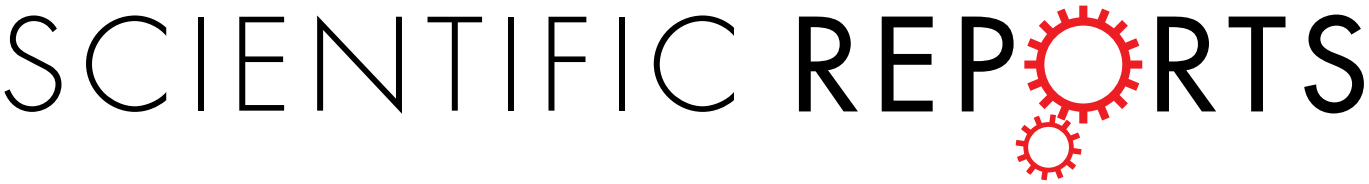

OPEN

Reliability in adolescent fMRI within two years – a comparison of three tasks

Received: 3 June 2016
Accepted: 11 April 2017
Published online: 23 May 2017

Nora C. Vetter[1,2,3], Julius Steding[1,2,4], Sarah Jurk[1], Stephan Ripke[1], Eva Mennigen[1] & Michael N. Smolka[1]

Longitudinal developmental fMRI studies just recently began to focus on within-subject reliability using the intraclass coefficient (ICC). It remains largely unclear which degree of reliability can be achieved in developmental studies and whether this depends on the type of task used. Therefore, we aimed to systematically investigate the reliability of three well-classified tasks: an emotional attention, a cognitive control, and an intertemporal choice paradigm. We hypothesized to find higher reliability in the cognitive task than in the emotional or reward-related task. 104 healthy mid-adolescents were scanned at age 14 and again at age 16 within M = 1.8 years using the same paradigms, scanner, and scanning protocols. Overall, we found both variability and stability (i.e. poor to excellent ICCs) depending largely on the region of interest (ROI) and task. Contrary to our hypothesis, whole brain reliability was fair for the cognitive control task but good for the emotional attention and intertemporal choice task. Subcortical ROIs (ventral striatum, amygdala) resulted in lower ICCs than visual ROIs. Current results add to the yet sparse overall ICC literature in both developing samples and adults. This study shows that analyses of stability, i.e. reliability, are helpful benchmarks for longitudinal studies and their implications for adolescent development.

To date, the field of longitudinal developmental fMRI studies is growing[1]. However, it remains largely unclear which degree of quantitative reliability can be achieved in developmental studies.

The preferable quantitative reliability measure in fMRI studies is the intraclass coefficient (ICC[2]) with the following formula:

$$\mathrm{ICC}(3,\,1) = \frac{MS_{between} - MS_{error}}{MS_{between} + (k-1)MS_{error}} \quad (1)$$

The total sum of squares in this model is split into between-subjects ($MS_{between}$) and error (MS_{error}) mean sums of squares and k represents the number of obvervations[3]. The ICC ranging from 0 to 1 tells us how much variance from the total variance in two measurements is due to variance between participants. An ICC of 1 would imply that participants' brain activation does not change over time (no within-subject variance). ICCs are classified according to Cicchetti[4] as poor (<0.40), fair (0.41–0.60), good (0.61–0.75), and excellent (>0.75)[5]. So far, almost exclusively adult neuroimaging studies measured reliability and found large variance across studies with an average ICC of 0.5[2]. According to Cicchetti[4], this ICC can be classified as 'fair'. These methodical studies measured only small samples of 10 to 20 adults in a short time span from a few days to a few weeks[2].

However, it remains largely unanswered if these test-retest reliabilities can be generalized to typical developmental longitudinal samples, which usually span larger time intervals between measurements. There have been only two previous developmental studies that reported ICCs[6,7]. Van den Bulk *et al.*[7] investigated n = 20 12 to 19 year-old adolescents and obtained fair reliability for the prefrontal cortex (PFC) and poor reliability for the amygdala using an emotional faces task. Koolschijn *et al.*[6] used a cognitive rule-switch task and showed fair to

[1]Department of Psychiatry and Neuroimaging Center, Technische Universität Dresden, Dresden, Germany. [2]Department of Child and Adolescent Psychiatry, Faculty of Medicine of the TU Dresden, Dresden, Germany. [3]Department of Psychology, Bergische Universität Wuppertal, Wuppertal, Germany. [4]Division of Psychological and Social Medicine and Developmental Neurosciences, Faculty of Medicine of the TU Dresden, Dresden, Germany. Correspondence and requests for materials should be addressed to N.C.V. (email: nora.vetter@tu-dresden.de) or M.N.S. (email: michael.smolka@tu-dresden.de)

good reliability for n = 12 15 year-old adolescents. The two studies of van den Bulk *et al.*[7] and Koolschijn *et al.*[6] differ in their investigated age-range, time interval (van den Bulk *et al.* 3 months; Koolschijn *et al.* 4 years) and their employed task domain that was either cognitive or emotional. Thus, evidence on reliability in developmental studies remains sparse. To fill this research gap the current study aimed at analyzing reliability in a large sample of 104 14-year old adolescents measured within a time interval of 2 years. Methodically, we focused on two important factors that can influence reliability: the task domain and the region of interest (ROI).

The task domain is a first factor that might influence fMRI reliability. Adult studies showed that reliabilities differed between task domains such as cognitive, emotional, or reward-related[2]. Only one adult study compared the reliability between these task domains using specific ROIs in one sample[8]. Results indicated a poor ICC for the amygdala in an emotional faces task, fair ICCs for frontal and parietal regions in a cognitive N-Back task, and fair to good ICCs in the ventral striatum (VS) for a reward task. Taken together, this study suggests that ICCs might be higher in cognitive and reward-related compared to emotional task domains.

Currently, there is no developmental reliability study comparing task domains. This is surprising since a recent review on developmental longitudinal studies suggests emotional and reward-related tasks might show lower test-retest reliability than cognitive tasks[1]. This was concluded from findings of low reliability, both for amygdala activity in emotional tasks[7,9,10] and VS activity in reward tasks[11,12]. In contrast, the prefrontal and parietal cortex showed relative high reliability in cognitive control tasks[6,13]. Most of these studies except Koolschijn *et al.*[6] and van den Bulk *et al.*[7], however, did not measure adolescent ICCs but either analyzed only Pearson's correlations of time point one and two[11,12], only reported on group differences of activation from time point one and two[9,10], or analyzed ICCs only in an adult sub-sample[13]. In contrast to Pearson's correlations the ICC provides a more accurate estimate because it can distinguish between systematic variation and average consistency over time[14]. Group differences are also not appropriate for conclusions about reliability because they only compare activation on a group level instead of an individual level. Therefore, the ICC is most suited as a quantitative intra-individual measure of reliability.

With this in mind, for the first time, we aimed at systematically comparing an emotional, a cognitive, and a reward-related task in an adolescent sample. The emotional task has been shown to yield valid results both on the behavioral and neural level[15,16]. It activates the fusiform gyrus, the inferior and middle frontal gyrus, and the inferior parietal lobe. Amygdala activation for negative stimuli in this task has been demonstrated to be sensitive towards a family history of depression in healthy adolescents[15]. The cognitive control task has shown robust switch and interference effects on the behavioral and on the neural level[17,18]. Further, the neural overlap between the switch and interference effect has revealed brain activation in the dorsal anterior cingulate cortex (dACC), the dorsolateral prefrontal cortex (dlPFC) as well as the posterior parietal cortex (PPC)[17]. The intertemporal choice task[19] is a widely used task that activates the VS for value processing and the ACC, PFC, and PPC for intertemporal decision making[20–22]. Developmental change in activation from age 14 to 16 has only been found for the emotional attention task[16], while the other tasks did not yield developmental effects[23,24].

A second factor influencing fMRI reliability is the chosen ROI. While developmental emotional tasks suggest lower reliability for the amygdala[1,7], higher reliabilities seem to result for occipital regions[7]. Previous studies mostly focused on only one or two regions such as the amygdala for emotional tasks[8,25,26]. Here, we analyzed three to five functional ROIs important for the respective task to achieve an overall picture of test-retest reliability. Additionally, we analyzed the whole brain ICC because it calculates the global concordance of neural activation regarding all voxels and therefore has been suggested to be the strictest criterion of fMRI reliability[2].

While considering the two important factors task domain and ROI, other parameters that might influence reliability[2,8] were held constant: scanner, scanning parameters, sample size, time interval, and event-related task design across all paradigms. We expected that the task of the cognitive domain would show higher reliability than that of the emotional or reward-related domain considering adult[8] and current developmental literature[1].

Results

Behavioral reliability. Behavioral reliability was fair in all behavioral measures of the three paradigms except for the overall reaction time of the cognitive control task, in which reliability was good (see Table 1). This fair to good reliability fits to the behavioral developmental effects in all paradigms: Adolescents became faster from age 14 to 16 in both the emotional attention[16] and cognitive control paradigm. The log-transformed discount parameter increased which can probably be interpreted with decreased impulsivity from age 14 to 16[22].

FMRI reliability. *Whole brain ICCs.* The whole brain ICC of the reward paradigm was highest across paradigms, 0.74 (see Fig. 1), and together with the emotional attention paradigm, 0.62 (see Fig. 2), it was in the "good" range. The ICC of the cognitive control paradigm was lower and only in the fair range, 0.44 (see Fig. 3). An ANOVA showed that the whole brain reliability differed significantly between the paradigms ($F = 102.67$, $p < 0.001$, $\eta^2_{partial} = 0.499$) with post-hoc analyses revealing that whole brain reliability of the reward paradigm was higher than emotional attention, which was higher than cognitive control (with all paradigms differing significantly from another, p's < 0.001).

ICCs of different ROIs. For the task-based ROIs in the emotional attention paradigm, ICCs were poor (amygdala, IFG, medial prefrontal cortex, mPFC) to excellent (fusiform gyrus, FG) ranging from 0.19 to 0.84 and poor for the development-based ROIs (ACC, IFG) ranging from 0.17 to 0.29 (see Fig. 2). A Wilcoxon signed-rank test revealed that the task-based IFG ROI was higher than the development-based IFG ROI ($p = 0.002$ for the left IFG and $p = 0.001$ for the right IFG). For the cognitive control paradigm, ICCs ranged from 0.32 to 0.56. ICCs were thus poor to fair for the dlPFC and dACC and fair for the PPC (see Fig. 3). The intertemporal choice paradigm yielded poor to fair ICCs for the VS and ACC ranging from 0.32 to 0.52 and excellent ICCs for the superior

Task	Behavioral Measure	T1 - ms M (SD)	T2 - ms M (SD)	t/p[a]	d[b]	$ICC_{(3,1)}$ (95%-CI)
Emotional attention	RT (overall)	719 (85)	696 (101)	2.33/0.022	0.24	0.46 (0.29–0.60)
	RT (negative attended)	726 (87)	700 (109)	2.45/0.016	0.26	0.42 (0.25–0.57)
Cognitive control	RT (overall)	906 (151)	826 (127)	7.15/<0.001	0.57	0.67 (0.55–0.76)
	RT (switch incongruent)	992 (162)	905 (142)	7.04/<0.001	0.57	0.46 (0.29–0.60)
Intertemporal choice	log_k[c]	−4.73 (0.79)	−4.93 (0.98)	2.18/0.032	0.22	0.47 (0.31–0.61)

Table 1. Behavioral data at both time points and resulting ICCs. Note: [a]t-test for paired samples comparing T1 and T2 values; [b]Cohen's d for the standardized mean difference; [c]log-transformed discount parameter, for methods, see Ripke *et al.*[22].

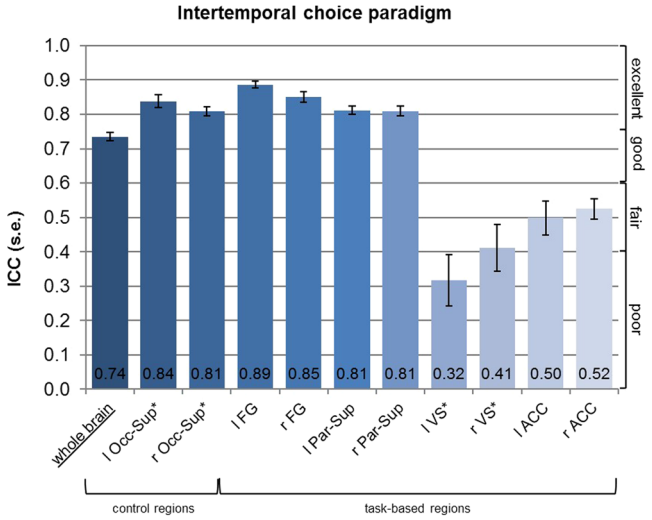

Figure 1. Results of ICC analyses for the intertemporal choice paradigm: *These regions are based on anatomical masks (AAL). l – left, r – right, Occ-Sup – Superior occipital lobe, FG – Fusiform gyrus, ACC – Anterior cingulate cortex, Par-Sup – Superior parietal lobe.

parietal lobe (Par-Sup) and the FG ranging from 0.81 to 0.89 (see Fig. 1). The control region in the occipital cortex (superior occipital lobe, Occ-Sup) yielded good to excellent reliability across paradigms.

Discussion

The current study aimed at investigating reliability in a large sample of mid-adolescents in three important domains of information processing using an emotional, a cognitive, and a reward-related task. We also considered different ROIs while holding other parameters that might influence reliability constant. Results showed that behavioral reliability was fair for all three paradigms. For fMRI reliability, the cognitive task yielded only fair whole brain reliability whereas the emotional and the reward-related task showed good whole brain reliability. ICCs of ROIs depended largely on the specific regions and the task and ranged from poor to excellent. Overall, ICCs were comparable to previous adult studies[2].

In terms of behavioral reliability from age 14 to 16, we found fair to good ICCs. For the emotional and the cognitive task, the participants became faster, while the reward-related task indicated less impulsive behavior, which is in line with previous studies[27–29]. Low behavioral reliability can be expected for tasks with developmental changes.

For our first aim, to explore whether functional imaging reliability depends on the task domain, the whole brain ICC was chosen. This ICC has been suggested to be the strictest approach for reliability[2], because it assumes on a whole-brain basis that the level of activity in all voxels should remain constant irrespective of suprathreshold activation. All three paradigms obtained a fair to good whole brain reliability. The reward paradigm had the highest whole brain ICC of 0.74 which can be classified as good to excellent. The emotional task had a whole brain ICC of 0.62 that was significantly lower but still in the good range. The ICC of the cognitive task differed significantly and was only in the fair range (ICC = 0.44). Thus, our first hypothesis, that the cognitive task would show higher reliability than the emotional or reward-related task (see also ref. 1) was not supported. To the contrary, the reward-related task yielded highest reliability followed by the emotional and the cognitive task. Our findings do not support the conclusions of Plichta *et al.*[8], probably because they measured adults and investigated the amygdala only for their emotional task while we also investigated whole brain ICCs. Speculatively for the cognitive paradigm, the low behavioral reliability might probably be related to the low fMRI reliability. However,

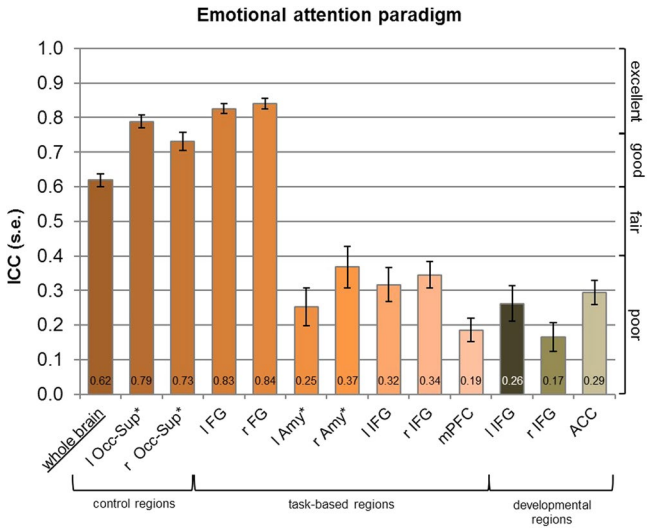

Figure 2. Results of ICC analyses for the emotional paradigm. *These regions are based on anatomical masks (AAL). l – left, r – right, Occ-Sup – Superior occipital lobe, FG – Fusiform gyrus, Amy –Amygdala, IFG – Inferior frontal gyrus, mPFC – Medial prefrontal cortex, ACC – Anterior cingulate cortex.

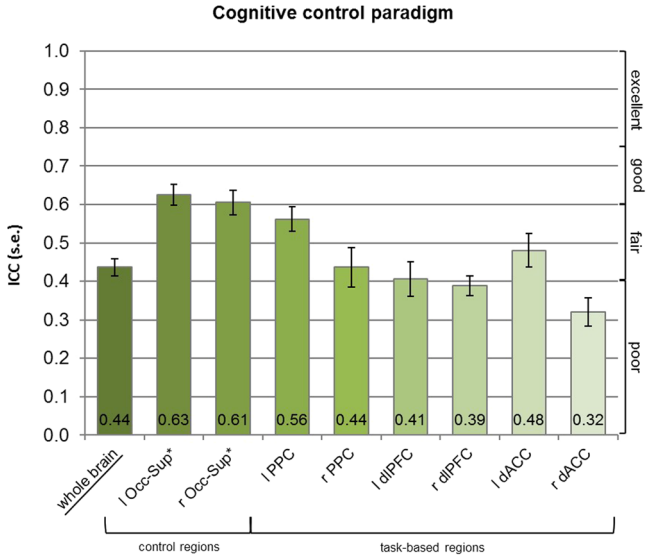

Figure 3. Results of ICC analyses for the cognitive control paradigm. *These regions are based on anatomical masks (AAL). l – left, r – right, Occ-Sup – Superior occipital lobe, PPC – Posterior parietal cortex, dlPFC – Dorsolateral prefrontal cortex, dACC – Dorsal anterior cingulate cortex.

when exploring correlations of behavioral and fMRI ICCs we did not find such a relationship (see S4 in the supplements). Furthermore, the lower whole brain ICCs of the cognitive control paradigm could stem from lower ICCs in the occipital regions (0.61 and 0.61 as opposed to 0.79, 0.73, 0.84 and 0.81 for the other two paradigms), respectively higher ICCs in the emotional attention paradigm for lower processing regions such as the IFG. The conclusion regarding lower reliability in the cognitive control paradigm has thus to be taken cautiously and investigated further in future studies.

Regarding our second aim, the single analyzed ROIs, the control region in the occipital cortex yielded good to excellent reliability across paradigms. The high reliability for the occipital cortex in the emotional paradigm is in line with another adolescent study[7]. The rather low-level visual area fusiform gyrus also yielded excellent reliability in both the emotional and reward task in line with previous emotional adult studies[30,31]. In contrast, other regions that are relevant for cognitive or emotional-motivational processes such as subcortical (amygdala, VS) and cortical regions (PFC) showed low reliability. Taken together, the current study suggests that across three tasks in the same sample reliabilities might be higher in regions of basic visual processing compared to cognitive or emotional-motivational brain regions. This might be due to higher variability in higher-level cognitive

processes than basic visual processing[32]. Another explanation might be that developmentally, visual regions have already matured, while subcortical and cortical higher-level regions continue to develop in adolescence[33,34].

In the following the regions that are relevant for cognitive or emotional-motivational processes are discussed for each paradigm separately.

For the *emotional attention paradigm* we found poor amygdala ICCs. Only one previous study investigated *adolescent* amygdala reliability with an age-heterogeneous sample of n = 20 12 to 19 year-olds[7] and found poor reliability within a short interval of 3 months. Our results show that poor amygdala reliability is also evident in a large sample of mid-adolescents within a longer time interval of 2 years.

From a developmental perspective, current results can be integrated with previous findings of a potential peak in amygdala activation in mid-adolescence compared to child- and adulthood (for a review, see refs 1 and 33). While some previous cross-sectional studies have supported this amygdala peak[35,36], longitudinal studies rather indicated "relative stability" in amygdala activation across mid-adolescence[9,16]. The current sample is a sub-sample of our previous longitudinal study that did not find amygdala activation change from age 14 to 16[16]. Therefore, current results suggest that this "relative stability" and lack of peak in mid-adolescence might occur at the same time as intra-individual variability, i.e. low reliability in amygdala activation (in accordance with the conclusions of a recent review)[1].

It is also possible that the amygdala signal itself might be instable, independent of development[1]. This is supported by adult studies that also found poor to fair amygdala ICCs in emotional tasks[8,25,26,30,31].

Regarding frontal regions important for emotional processing[16], the first region IFG showed poor reliability in line with a previous emotional adult study[31]. An emotional adolescent study found that IFG activation at baseline correlated with activation 2 years later indicating some degree of reliability[37]. The second region, mPFC, showed poor reliability similar to the adolescent study of van den Bulk *et al.*[7]. In our previous longitudinal study[16], part of the IFG and the ACC demonstrated a developmental effect, i.e. higher activation at age 16 than 14. Expectedly, this developmental region showed a lower reliability than the (larger) IFG ROI that was functionally defined at age 14. The ACC showed a poor reliability similar to an adult study[31].

The *cognitive control paradigm* showed poor to fair ICCs partly in contrast to the only other adolescent study[6] that found good ICCs for the PPC and dACC while the dlPFC result was in a similar fair range. But it should be noted, that the ACC of Koolschjin *et al.*[6] was located more anteriorly. Also an adult study found good to excellent ICCs[38]. However, there are not many studies that have calculated ICCs in cognitive control tasks. Cognitive control can be divided into three related factors: inhibition, shifting, and updating[39]. The current interference *and* switch task assesses both inhibition and shifting. No previous study examined ICCs using such a task. Taking updating tasks into account, current results are in line with ICC ranges of adult studies (Plichta *et al.*[8] using an n-back task, Brandt *et al.*[40] using a memory encoding task, and Bennett and Miller[41] using an episodic and two-back memory task). We speculate that ICCs in our task may be low, as it assesses two cognitive control functions simultaneously. Unfortunately, due to our task design it is not possible to separate both components of cognitive control (i.e. task switching and overcoming incongruence) because each trial contains information on incongruence as well as task switching. Future studies should systematically compare ICCs of more basic cognitive control tasks.

To our knowledge this is the first study that tested reliability of a *reward-related paradigm* in an adolescent sample. The intertemporal choice paradigm showed fair to good ICCs in the superior parietal lobe and the ACC, which is in line with previous adult studies (probabilistic reversal task[42]; classification learning task[43]). For the VS, our results were in the poor to fair range, which is in line with Chase and colleagues[44] using a card guessing task re-scanned within one week. In contrast, Plichta *et al.*[8] found excellent ICCs in the VS for a reward task within two weeks. Our findings of low VS reliability are in line with the conclusions by Crone & Elzinga[1] that there might be large variability in subcortical brain regions (amygdala, VS) in adolescence.

The reliability of fMRI data has implications for longitudinal studies of reward processing, which are pivotal to detect developmental change in brain-behavior relations. For example, Braams *et al.*[45] assessed response to rewards in participants aged 8 to 25 longitudinally within 2 years and found an inverted U-shaped activation of the VS with a peak in activation during adolescence. This peak was also found behaviorally in a balloon analog risk taking task. A further longitudinal study was able to extent knowledge about dynamics of reward anticipation on the brain and behavioral level in adolescents[11]. Results showed that changes in VS activation over 2 years were related to changes in the behavioral approach system fun seeking score[46] during the same time period. A third longitudinal study found increasing dorsal striatal activation from mid-adolescence to late-adolescence/early adulthood in response to anticipation of gain and loss[12]. Taken together, reliability of reward-related activation seems to depend on time between measurements and brain regions. While ICCs of cortical areas were mostly good to excellent, the results regarding the subcortical area VS are not conclusive. Additionally, ICCs have to be interpreted with respect to expected developmental-related changes regarding activation patterns. Thus, additional studies are needed to systematically investigate this relationship.

Overall, current results warrant discussion with regard to the following considerations and limitations. The ICC depends on the between-subject variance. Thus, current results might be related to the type of the current sample that is rather homogenous (fine-grained age range, similar sociodemography, intelligence, and pubertal status). Future studies could test reliability using more heterogenous samples.

Similar to other adolescent reliability studies[6] this study was not designed a priori as a methodological study that investigates reliability but part of an overall research project focusing on adolescent brain development in several domains. The large sample size spanning about 200 participants (before exclusion due to movement, technical or behavioral outliers, see S1 in the supplement) required a time span of about 2 years. Because of this time span and the developmental sample we can therefore not disentangle between reliability due to development or reliability which would have occurred without development (e.g. in an adult population).

Assuming that changes in brain processes will be more likely to occur in contrasts which are expected to be effected by development (i.e. specific contrasts, like decision for small immediate vs. larger later in the intertemporal choice task), we used more general contrasts to investigate the reliability of the imaging data in our large sample. Although reliability and developmental changes are not two sides of the same coin, both are harder to distinguish the more developmentally sensitive the contrast is. Therefore, our rational was that, if the reliability of the more general contrasts would be moderate to high, the imaging data per se might be reliable; in the current study even over a timespan of two years.

As this area is still controversial, we chose baseline contrasts after careful consideration, since their constancy allowed us to compare single conditions of different paradigms more clearly as opposed to two contrasted conditions per paradigm. Especially in the developmental literature, the importance of differentiating between baseline and higher level contrasts has been emphasized[1,47] to infer more precisely which contrast led to developmental effects: in case of developmental changes in a higher level contrast, it is not possible to conclude what has changed: condition A, condition B, or both[1,47]. Furthermore, it has been suggested that baseline contrasts yield better reliability than higher level contrasts[8]. However, current results have to be considered carefully and with potentially lower ICCs for higher level contrasts in mind.

Nevertheless, the study is unique due to its large sample and the three tasks that were tested for reliability. Future studies could systematically assess reliability in a (smaller) adolescent sample within a short time span and at the same time systematically control for potential changes in several domains (development, cognitive strategy, motivation etc.) and compare tasks that show developmental change in adolescence and those which do not. The reliabilities could further be compared to an additional adult population.

This study contained a qualitative comparison between tasks and was not designed a priori to systematically compare reliabilities of parallelized tasks. There were several aspects that could not be controlled for in the current analyses. First, the number of specific trials for the chosen contrast differed between tasks. While the task with the highest amount of trials was the most reliable one, the emotional attention task had fewer trials than the cognitive control task but a higher reliability, which might not fit to the conclusion that amount of trials correlates with task reliability. Second, behavioral differences that might stem from changes in performance, cognitive strategy or task focus[48–50] could not be controlled for. Third, the implicit baseline that was included in all regressors of interest differed between tasks (length of fixation cross and cognitive process during baseline). Fourth, due to each paradigm's specific effect size functional ROIs were created specifically for each paradigm: the statistical thresholds for the second-level analyses that built the basis of the definition of the functional ROIs differed between paradigms as well as the approach to rely on the peak voxels (emotional attention, cognitive control) or the anatomical overlap (intertemporal choice). Future studies should hold these features between tasks constant or control for them to be able to systematically compare task domains without potential confounders. ROIs were defined on the group level instead of the individual level similar to other studies[20,51–55]. Future studies could also add ROIs based on the individual level and calculate reliability.

Taken together, ICCs in each paradigm were largely dependent on the respective ROIs with subcortical ROIs (VS, amygdala) resulting in lower ICCs than visual ROIs. The emotional and reward paradigm had higher whole brain ICCs than the cognitive paradigm. Current results add to the yet sparse overall ICC literature in both developing samples and adults. In the different task domains, ICCs were similar as in adult studies. To test whether results are specific for adolescents or can be generalized to adults the current paradigms could be tested in adults. Analyses of stability, i.e. reliability, are helpful benchmarks for longitudinal studies and their implications for adolescent development.

Material and Methods

Participants. The institutional review board of the medical faculty of the TU Dresden approved the study and the study was realized in accordance with it and with the Declaration of Helsinki. Participants were recruited from local schools and received monetary compensation for their participation. Written informed consent was obtained from both the participants and one of their legal guardians. The current dataset stems from the overall project "The adolescent brain"[22] that investigated 250 adolescents at age 14 and again at age 16. For technical and practical issues not all of these participants completed all three tasks at both time points.

Sub-populations of this sample were previously reported regarding cross-sectional analyses of age 14 (emotional attention task, n = 164, Pilhatsch *et al.*[15], intertemporal choice task, n = 235, Ripke *et al.*[22]; n = 206, Ripke *et al.*[56], cognitive control task, n = 184, Mennigen *et al.*[17], Rodehacke *et al.*[18]) or longitudinal change from age 14 to 16 (emotional attention task, n = 144, Vetter *et al.*[16], intertemporal choice task, n = 80, Ripke *et al.*[23]). We here report on the overlapping sample of 104 healthy participants who performed all three tasks at age 14 and 16 successfully. This sample was analyzed for reliability for the first time.

For information of exclusion criteria for each task see Supplement S1. Participants had normal or corrected to normal vision and neither any record nor any current diagnoses of neurological, psychiatric, or serious medical disorders. Current psychiatric disorders were identified with the Development and Well-Being Assessment (DAWBA[57]). General cognitive ability of the sample was in the average to above average range (IQ across both time points: $M = 115$; $SD = 10$; range = 89–139) and did not change between measurements ($t = 1.03$; $p = 0.31$). 76.7% of the participants were visiting the higher grammar school (German "Gymnasium") and 23.3% the lower grammar school (German "Mittelschule"). Parental education ranged from no school education (7) to doctoral degree (1) with an average education of $M = 3.38$ ($SD = 1.45$), representing a university diploma. For further details about the sample see Table 2. A urine test assured no use of illicit drugs (e.g. cannabis, heroin, cocaine) at the day of assessment.

Age in years at T1	M = 14.52, SD = 0.32, range 13.83–14.99
Age in years at T2	M = 16.55, SD = 0.34, range 15.86–17.21
Interscan interval in years	M = 2.03, SD = 0.11, range 1.84–2.38
No. of females	N = 54 (51.9%)
No. of right-handers	93 (1 bimanual, 10 left)
IQ at T1[a]	M = 114, SD = 10, range 86–135
IQ at T2[b]	M = 115, SD = 11, range 91–145
Pubertal status[c] at T1	M = 3.65, SD = 0.65, i.e. mid- to late pubertal status
Pubertal status at T2	M = 4.18, SD = 0.57, i.e. late pubertal status

Table 2. Participant characteristics (n = 104). *Note*. [a]measured with the Wechsler Intelligence Scale For Children (WISC) that consisted of the subtests Similarities, Block Design, Vocabulary, and Matrices[61]; [b]measured with the Wechsler Adult Intelligence Scale (WAIS) that consisted of the same subtests as WISC and additionally the Letter-Number Sequencing, Symbol Search, Digit Span, and Coding[62]; [c]Pubertal status ranges from 1 for prepubertal to 5 for postpubertal status, measured with the Pubertal Development Scale (PDS[63]).

	emotional attention	**cognitive control**	**intertemporal choice**
No. of trials of the chosen contrast/total task trials	20/120	64/256	90/90
Duration in min	15	21	25
Regressors of interest	negative attended > implicit baseline	switch incongruent > implicit baseline	intertemporal decision phase > implicit baseline
Task design	event-related	event-related	event-related
Regions of interest			
Task-based	mPCF	dACC	ACC
	IFG	dlPFC	Par-Sup
	Amy	PPC	VS
	FG		FG
Developmental	IFG	none	none
	ACC		
control region	Sup-Occ	Sup-Occ	Sup-Occ

Table 3. Overview of task characteristics. *Note*. mPFC – medial prefrontal cortex, IFG – inferior frontal gyrus, Amy – Amygdala, FG – fusiform gyrus, ACC – anterior cingulate cortex, Sup-Occ – superior occipital lobe, dACC – dorsal anterior cingulate cortex, dlPFC – dorsolateral prefrontal cortex, PPC – posterior parietal cortex, Par-Sup – superior parietal lobe, VS – ventral striatum.

Paradigms. For an overview of the main characteristics of the three paradigms see Table 3. In the emotional attention task, participants had to decide whether a pair of visual target stimuli was identical or not while another pair was presented as a distractor. Participants were not asked to attend to a particular emotional category but cued spatially by an arrow pointing in the direction of the two stimuli. Each trial consisted of a pair of pictures from one of three emotional categories (positive, neutral, negative) and a pair of non-emotional pictures. The emotional pictures were taken from the International Affective Picture System (IAPS[58]); and the non-emotional pictures were created by shredding the chosen IAPS pictures with GIMP (www.gimp.org). For further details see Vetter *et al.*[16] and Pilhatsch *et al.*[15] and Supplement S2.

The first screen of the cognitive control task was an arrow consisting of two triangles pointing in one (left, right, up or down) direction and a red dot located either at the tip or the tail of the arrow. Participants were instructed to move a joystick in the direction indicated by the arrow or the dot. The shape of the background served as a task cue: If the background was rectangular, participants had to move the joystick in the direction of the arrow and ignore the position of the dot; conversely, if the background was circular, participants had to respond to the position of the dot while ignoring the arrow direction. Stimuli could be congruent, i.e. dot and arrow were pointing in the same direction, or incongruent, i.e. the dot and the arrow were pointing in opposite directions. For further details see Mennigen *et al.*[17], Rodehacke *et al.*[18].

In the intertemporal choice task participants had to choose between a larger later reward, which changed from trial to trial and a fixed immediate reward, which was instructed beforehand but not shown during scanning. In the current paper, the contrast of interest was the phase of the presentation of the potential later reward, i.e. the intertemporal decision phase, which refers to the process of comparing both alternatives in a given trial (fixed immediate or later reward). The task started with a behavioral training session to estimate the individual

impulsivity parameter *k*, which was used to adapt the scanning paradigm to the subjects' impulsivity. For more details see Ripke *et al.*[22] and Ripke *et al.*[56].

Task presentation and order. The paradigms were presented with a LCD-based display system which was mounted on the head-coil (NordicNeuroLab AS, Bergen, Norway). Behavioral data were collected with a joystick (Resonance Technology Inc., Northridge, CA, USA) for the cognitive control task and by ResponseGrips (©NordicNeuroLab) with a button on a grip in each hand for the emotional attention and intertemporal choice task. Task presentation and recording of the behavioral responses was performed using Presentation® software (version 11.1, Neurobehavioral Systems, Inc., Albany, CA). Each task was preceded by a practice session. Since the tasks were assessed within an overall project including a large behavioral and fMRI battery, the order of tasks varied slightly between time points. At age 14, the order of paradigms was emotional attention, cognitive control and intertemporal choice on three different days within two weeks. At age 16 first the cognitive control and then the intertemporal choice task were assessed on the same day followed by the assessment of the emotional attention task within two weeks.

Functional imaging. *Image acquisition.* For all three paradigms and across both sessions, image acquisition remained the same. MRI data was acquired using a 3 T whole-body MR tomograph (Magnetom TRIO, Siemens, Erlangen, Germany) with a 12-channel head coil. For all paradigms and across both sessions, an identical standard Echo Planar Imaging (EPI) sequence was used for functional imaging (TR/TE: 2410/25 ms; flip angle: 80°). FMRI scans were obtained from 42 transversal slices. Voxel size was $3 \times 3 \times 3$ mm (slice thickness: 2 mm with 1 mm gap; FOV: 192×192 mm; in-plane resolution 64×64 pixels). Furthermore, a 3D T1-weighted magnetization-prepared rapid gradient echo (MPRAGE) image data set was acquired (TR/TE: 1900/2.26 ms; FOV: 256×256 mm; 176 slices; $1 \times 1 \times 1$ mm voxel size; flip angle: 9°). Scanning settings and protocols were identical for all three paradigms and across both time points.

Analysis of fMRI data. FMRI data analyses were performed using SPM5 (Wellcome Trust Center of Neuroimaging, London, UK) and were the same for both time points per paradigm.

Preprocessing. For preprocessing, which was identical for all three tasks, functional images were first slice-time corrected by using the middle slice as reference and realigned to the first image (by 6° rigid spatial transformation). Afterwards they were spatially normalized into Montreal Neurological Institute (MNI) space and spatially smoothed with an 8 mm full-width half maximum Gaussian kernel.

Statistical analysis. For all paradigms first-level contrasts were computed with a fixed effects analysis for each participant based on the general linear model by modeling the different conditions as regressors of interest within each voxel for the whole brain. For each paradigm, the six subject-specific movement regressors, which were derived from the rigid-body realignment, were included as covariates of no interest. A high-pass filter with cut-off 128 s was applied to remove the low frequency physiological noise[59] for each paradigm. Also an autoregression, AR(1), model was employed for the residual temporal autocorrelation[59] for each paradigm. Contrasts of interest (see Table 3) were computed for each paradigm within each subject. The first-level contrast images from the weighted beta-images were used for second-level whole brain random-effects analyses to allow for population inference. For a detailed description of the first- and second-level analyses for each paradigm see S3 in the supplement.

Definition of ROIs. For an overview of used ROIs see Fig. 4. ROIs were defined based on a priori hypotheses regarding activation in the respective tasks and based on functional masks resulting from the whole-brain analyses of each task at the first time point, i.e. age 14[16,17,22]. 10 mm spheres were placed around the peak coordinates (see Table S3 in the Supplementary Materials) of the whole brain analyses at age 14 and thus final ROIs created. Additionally, bilateral superior occipital ROIs using the WFU-PickAtlas with the Automated Anatomical Labeling Atlas (AAL) were created that served as control regions for all three tasks. Specific ROI approaches for each paradigm are described in the following.

Emotional attention paradigm. For this paradigm, we focused on attending negative versus attending neutral stimuli for functional ROI extraction for two reasons: The attending negative in contrast to the attending neutral condition resulted in slower reaction times which indicates an attentional capture effect[16]. Second, separate ROIs for emotional attention could be created by subtracting the neutral contrast (but not by subtracting the implicit baseline since almost the whole brain was activated). The amygdala was chosen as an additional ROI because it was also activated for negative target stimuli in the paradigm but defined the whole amygdala as a larger cluster anatomically using the WFU-PickAtlas with the Talairach Daemon (TD) Brodman atlas (following[15,16]). Furthermore, for this paradigm, two ROIs with developmental effects were analyzed that emerged from higher activation during presentation of emotional target and distractor stimuli for age 16 versus 14 in the right and left inferior frontal gyrus (IFG) and the ACC[16], see Table S3 in the Supplementary Materials.

Cognitive control paradigm. ROIs were created based on a conjunction analysis[17]. Switch- and incongruence-related activity overlapped in bilateral dACC, dlPFC and PPC. We chose trials with co-occurrence of incongruence and switch (switch incongruent trials > implicit baseline) because of two reasons. These trials led to a steep increase in reaction time and error rate therefore reflecting a high level of cognitive control[17]. Further, task switch and incongruence trials robustly and independently activated the core regions of the cognitive control network[17].

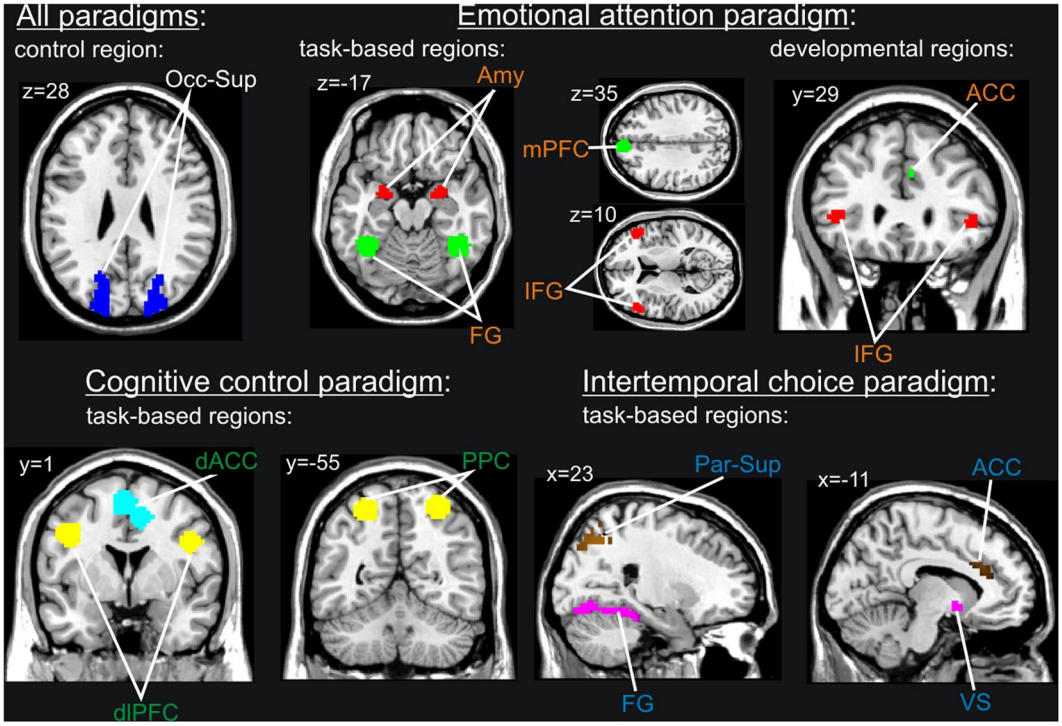

Figure 4. Regions of interest that were used to calculate ICC for all paradigms. The control regions for all paradigms were the left and right superior occipital lobe.

Intertemporal choice paradigm. For this paradigm, ROIs of the fusiform gyrus, the superior parietal lobe as well as the ACC were created by using the overlap of functional activation of the intertemporal decision phase[22,56] and anatomical regions using the WFU-PickAtlas with the AAL atlas. The overlap with anatomical regions was necessary to create distinct ROIs because the activation spanned one very large cluster across the whole brain. We additionally chose the VS as a ROI since it is highly relevant for reward paradigms. The anatomical ROIs of the VS were created with the WFU-PickAtlas using the AAL atlas.

Analyses of reliability. *Behavioral reliability.* Behavioral $ICCs_{(3,1)}$ were calculated using SPSS v21 (IBM Corp., Armonk, USA). For the emotional attention and the cognitive control paradigm, reaction times of the specific conditions and overall reaction times across conditions and for the intertemporal choice paradigm, log-transformed discount parameters were analyzed for reliability.

FMRI reliability. FMRI ICCs were calculated with the ICC toolbox of Caceres *et al.*[60]. We used the intra-voxel reliability "ICC_v" obtained by using the contrast value of each voxel within each ROI of each individual subject. The population estimate was obtained by bootstrapping with 1,000 re-samples of participants, of which medians and standard errors are reported. Additionally, whole brain ICCs were calculated, since this is the strictest criterion and potentially the most valuable reliability measure as it yields a global measurement of test-retest agreement[2]. ICCs were classified according to Cicchetti[4] as poor, <0.40, fair, 0.41–0.60, good, 0.61–0.75, and excellent, >0.75 (see also ref. 5).

References

1. Crone, E. A. & Elzinga, B. M. Changing brains: how longitudinal functional magnetic resonance imaging studies can inform us about cognitive and social-affective growth trajectories. *Wiley Interdiscip. Rev. Cogn. Sci.* **6**, 53–63, doi:10.1002/wcs.1327 (2015).
2. Bennett, C. M. & Miller, M. B. How reliable are the results from functional magnetic resonance imaging? *Ann. N. Y. Acad. Sci.* **1191**, 133–155, doi:10.1111/nyas.2010.1191.issue-1 (2010).
3. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **86**, 420–428, doi:10.1037/0033-2909.86.2.420 (1979).
4. Cicchetti, D. V. 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, doi:10.1076/jcen.23.5.695.1249 (2001).
5. Cicchetti, D. V. & Sparrow, S. A. Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *Am. J. Ment. Defic.* **86**, 127–137 (1981).
6. Koolschijn, P. C. M. P., Schel, M. A., de Rooij, M., Rombouts, Sa. R. B. & Crone, E. A. A three-year longitudinal functional magnetic resonance imaging study of performance monitoring and test-retest reliability from childhood to early adulthood. *J. Neurosci.* **31**, 4204–4212, doi:10.1523/JNEUROSCI.6415-10.2011 (2011).
7. van den Bulk, B. G. *et al.* How stable is activation in the amygdala and prefrontal cortex in adolescence? A study of emotional face processing across three measurements. *Dev. Cogn. Neurosci.* **4**, 65–76, doi:10.1016/j.dcn.2012.09.005 (2013).
8. Plichta, M. M. *et al.* Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. *Neuroimage* **60**, 1746–1758, doi:10.1016/j.neuroimage.2012.01.129 (2012).

9. Pfeifer, J. H. *et al*. Entering Adolescence: Resistance to Peer Influence, Risky Behavior, and Neural Changes in Emotion Reactivity. *Neuron* **69**, 1029–1036, doi:10.1016/j.neuron.2011.02.019 (2011).
10. Spielberg, J. M., Olino, T. M., Forbes, E. E. & Dahl, R. E. Exciting fear in adolescence: Does pubertal development alter threat processing? *Dev. Cogn. Neurosci.* **8**, 86–95, doi:10.1016/j.dcn.2014.01.004 (2014).
11. van Duijvenvoorde, A. C. K. *et al*. A cross-sectional and longitudinal analysis of reward-related brain activation: Effects of age, pubertal stage, and reward sensitivity. *Brain Cogn.* **89**, 3–14, doi:10.1016/j.bandc.2013.10.005 (2014).
12. Lamm, C. *et al*. Longitudinal study of striatal activation to reward and loss anticipation from mid-adolescence into late adolescence/early adulthood. *Brain Cogn.* **89**, 51–60, doi:10.1016/j.bandc.2013.12.003 (2014).
13. Ordaz, S. J., Foran, W., Velanova, K. & Luna, B. Longitudinal growth curves of brain function underlying inhibitory control through adolescence. *J. Neurosci.* **33**, 18109–24, doi:10.1523/JNEUROSCI.1741-13.2013 (2013).
14. Hunt, R. J. Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability. *J. Dent. Res.* **65**, 128–30, doi:10.1177/00220345860650020701 (1986).
15. Pilhatsch, M. *et al*. Amygdala-function perturbations in healthy mid-adolescents with familial liability for depression. *J. Am. Acad. Child Adolesc. Psychiatry* **53**, 559–568, doi:10.1016/j.jaac.2014.02.010 (2014).
16. Vetter, N. C., Pilhatsch, M., Weigelt, S., Ripke, S. & Smolka, M. N. Mid-adolescent neurocognitive development of ignoring and attending emotional stimuli. *Dev. Cogn. Neurosci* **14**, 23–31, doi:10.1016/j.dcn.2015.05.001 (2015).
17. Mennigen, E. *et al*. Exploring adolescent cognitive control in a combined interference switching task. *Neuropsychologia* **61**, 175–189, doi:10.1016/j.neuropsychologia.2014.06.022 (2014).
18. Rodehacke, S. *et al*. Interindividual differences in mid-adolescents in error monitoring and post-error adjustment. *PLoS One* **9**, 1–12, doi:10.1371/journal.pone.0088957 (2014).
19. Ainslie, G. Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychol. Bull.* **82**, 463–496, doi:10.1037/h0076860 (1975).
20. McClure, S. M., Laibson, D. I., Loewenstein, G. & Cohen, J. D. Separate Neural Systems Value Immediate and Delayed Monetary Rewards. *Science* **306**, 503–507, doi:10.1126/science.1100907 (2004).
21. Kable, J. W. & Glimcher, P. W. The neural correlates of subjective value during intertemporal choice. *Nat. Neurosci.* **10**, 1625–1633, doi:10.1038/nn2007 (2007).
22. Ripke, S. *et al*. Reward processing and intertemporal decision making in adults and adolescents: The role of impulsivity and decision consistency. *Brain Res.* **1478**, 36–47, doi:10.1016/j.brainres.2012.08.034 (2012).
23. Ripke, S., Mennigen, E., Müller, K. & Smolka, M. N. Who becomes impulsive: A longitudinal fMRI-Study of Inter-temporal Decision Making. In *HBM Congress* (2012).
24. Mennigen, E., Ripke, S. & Smolka, M. N. Influence of smoking on reward processing during inter-temporal choice: A longitudinal fMRI study. In *HBM Congress* (2014).
25. Johnstone, T. *et al*. Stability of amygdala BOLD response to fearful faces over multiple scan sessions. *Neuroimage* **25**, 1112–1123, doi:10.1016/j.neuroimage.2004.12.016 (2005).
26. Manuck, S. B., Brown, S. M., Forbes, E. E. & Hariri, A. R. Temporal Stability of Individual Differences in Amygdala Reactivity. *Am. J. Psychiatry* **164**, 1613–1614, doi:10.1176/appi.ajp.2007.07040609 (2007).
27. Cohen-Gilbert, J. E. & Thomas, K. M. Inhibitory Control During Emotional Distraction Across Adolescence and Early Adulthood. *Child Dev.* **84**, 1954–1966, doi:10.1111/cdev.2013.84.issue-6 (2013).
28. Fitzgerald, K. D. *et al*. The development of performance-monitoring function in the posterior medial frontal cortex. *Neuroimage* **49**, 3463–3473, doi:10.1016/j.neuroimage.2009.11.004 (2010).
29. Green, L., Fry, A. F. & Myerson, J. Discounting Of Delayed Rewards: A Life-Span Comparison. *Psychol. Sci.* **5**, 33–36, doi:10.1111/j.1467-9280.1994.tb00610.x (1994).
30. Sauder, C. L., Hajcak, G., Angstadt, M. & Phan, K. L. Test-retest reliability of amygdala response to emotional faces. *Psychophysiology* **50**, 1147–1156, doi:10.1111/psyp.12129 (2013).
31. Gee, D. G. *et al*. Reliability of an fMRI paradigm for emotional processing in a multisite longitudinal study. *Hum. Brain Mapp.* **36**, 2558–2579, doi:10.1002/hbm.22791 (2015).
32. Gorgolewski, K. J., Storkey, A. J., Bastin, M. E., Whittle, I. & Pernet, C. Single subject fMRI test-retest reliability metrics and confounding factors. *Neuroimage* **69**, 231–243, doi:10.1016/j.neuroimage.2012.10.085 (2013).
33. Casey, B. J., Jones, R. M. & Hare, T. A. The Adolescent Brain. *Ann. NY Acad. Sci.* **1124**, 111–126, doi:10.1196/annals.1440.010 (2008).
34. Giedd, J. N. *et al*. Brain development during childhood and adolescence: a longitudinal MRI study. *Nat. Neurosci.* **2**, 861–863, doi:10.1038/13158 (1999).
35. Hare, T. A. *et al*. Biological Substrates of Emotional Reactivity and Regulation in Adolescence During an Emotional Go-Nogo Task. *Biol. Psychiatry* **63**, 927–934, doi:10.1016/j.biopsych.2008.03.015 (2008).
36. Guyer, A. E. *et al*. A Developmental Examination of Amygdala Response to Facial Expressions. *J. Cogn. Neurosci.* **20**, 1565–1582, doi:10.1162/jocn.2008.20114 (2008).
37. Overgaauw, S., van Duijvenvoorde, A. C. K., Gunther Moor, B. & Crone, Ea A longitudinal analysis of neural regions involved in reading the mind in the eyes. *Soc. Cogn. Affect. Neurosci.* **10**, 619–627, doi:10.1093/scan/nsu095 (2015).
38. Sheu, L. K., Jennings, J. R. & Gianaros, P. J. Test-retest reliability of an fMRI paradigm for studies of cardiovascular reactivity. *Psychophysiology* **49**, 873–884, doi:10.1111/psyp.2012.49.issue-7 (2012).
39. 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, doi:10.1006/cogp.1999.0734 (2000).
40. Brandt, D. J. *et al*. Test-retest reliability of fMRI brain activity during memory encoding. *Front. Psychiatry* **4**, 1–9, doi:10.3389/fpsyt.2013.00163 (2013).
41. Bennett, C. M. & Miller, M. B. fMRI reliability: influences of task and experimental design. *Cogn. Affect. Behav. Neurosci.* **13**, 690–702, doi:10.3758/s13415-013-0195-1 (2013).
42. Freyer, T. *et al*. Test-retest reliability of event-related functional MRI in a probabilistic reversal learning task. *Psychiatry Res.* **174**, 40–46, doi:10.1016/j.pscychresns.2009.03.003 (2009).
43. Aron, A. R., Gluck, M. A. & Poldrack, R. A. Long-term test-retest reliability of functional MRI in a classification learning task. *Neuroimage* **29**, 1000–1006, doi:10.1016/j.neuroimage.2005.08.010 (2006).
44. Chase, H. W. *et al*. Accounting for Dynamic Fluctuations across Time when Examining fMRI Test-Retest Reliability: Analysis of a Reward Paradigm in the EMBARC Study. *PLoS One* **10**, e0126326, doi:10.1371/journal.pone.0126326 (2015).
45. Braams, B. R., van Duijvenvoorde, A. C. K., Peper, J. S. & Crone, E. A. Longitudinal Changes in Adolescent Risk-Taking: A Comprehensive Study of Neural Responses to Rewards, Pubertal Development, and Risk-Taking Behavior. *J. Neurosci.* **35**, 7226–7238, doi:10.1523/JNEUROSCI.4764-14.2015 (2015).
46. Carver, C. S. & White, T. L. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *J. Pers. Soc. Psychol.* **67**, 319–333, doi:10.1037/0022-3514.67.2.319 (1994).
47. Church, Ja, Petersen, S. E. & Schlaggar, B. L. The 'Task B problem' and other considerations in developmental functional neuroimaging. *Hum. Brain Mapp.* **31**, 852–862, doi:10.1002/hbm.21036 (2010).
48. Krishnan, S., Leech, R., Mercure, E., Lloyd-Fox, S. & Dick, F. Convergent and Divergent fMRI Responses in Children and Adults to Increasing Language Production Demands. *Cereb. Cortex* **25**, 3261–3277, doi:10.1093/cercor/bhu120 (2015).

49. Cohen Kadosh, K., Johnson, M. H., Dick, F., Cohen Kadosh, R. & Blakemore, S.-J. Effects of Age, Task Performance, and Structural Brain Development on Face Processing. *Cereb. Cortex* **23**, 1630–1642, doi:10.1093/cercor/bhs150 (2013).
50. Qin, S. *et al*. Hippocampal-neocortical functional reorganization underlies children's cognitive development. *Nat. Neurosci.* **17**, 1263–1269, doi:10.1038/nn.3788 (2014).
51. Bishop, S. J., Duncan, J. & Lawrence, A. D. State Anxiety Modulation of the Amygdala Response to Unattended Threat-Related Stimuli. *J. Neurosci.* **24**, 10364–10368, doi:10.1523/JNEUROSCI.2550-04.2004 (2004).
52. Peters, J. & Büchel, C. Episodic Future Thinking Reduces Reward Delay Discounting through an Enhancement of Prefrontal-Mediotemporal Interactions. *Neuron* **66**, 138–148, doi:10.1016/j.neuron.2010.03.026 (2010).
53. Nee, D. E., Wager, T. D. & Jonides, J. Interference resolution: Insights from a meta-analysis of neuroimaging tasks. *Cogn. Affect. Behav. Neurosci.* **7**, 1–17, doi:10.3758/CABN.7.1.1 (2007).
54. Hedden, T. & Gabrieli, J. D. E. Shared and selective neural correlates of inhibition, facilitation, and shifting processes during executive control. *Neuroimage* **51**, 421–431, doi:10.1016/j.neuroimage.2010.01.089 (2010).
55. Vuilleumier, P., Armony, J. L., Driver, J. & Dolan, R. J. Effects of Attention and Emotion on Face Processing in the Human Brain: An Event-Related fMRI Study. *Neuron* **30**, 829–841, doi:10.1016/S0896-6273(01)00328-2 (2001).
56. Ripke, S. *et al*. Common Neural Correlates of Intertemporal Choices and Intelligence in Adolescents. *J. Cogn. Neurosci.* **27**, 387–399, doi:10.1162/jocn_a_00698 (2015).
57. Goodman, R., Ford, T., Richards, H., Gatward, R. & Meltzer, H. The Development and Well-Being Assessment: Description and Initial Validation of an Integrated Assessment of Child and Adolescent Psychopathology. *J. Child Psychol. Psychiatry* **41**, 645–655, doi:10.1111/j.1469-7610.2000.tb02345.x (2000).
58. Lang, P. J., Bradley, M. M. & Cuthbert, B. N. International affective picture system (IAPS): Affective ratings of pictures and instruction manual (Tech. Rep. No. A-8) (2008).
59. Henson, R. In *Human Brain Functions* (eds Frackowiak, R. S. J., Friston, K. J. & Frith, C.) 793–822 (Elsevier Books, 2006).
60. Caceres, A., Hall, D. L., Zelaya, F. O., Williams, S. C. R. & Mehta, M. A. Measuring fMRI reliability with the intra-class correlation coefficient. *Neuroimage* **45**, 758–768, doi:10.1016/j.neuroimage.2008.12.035 (2009).
61. Petermann, F. & Petermann, U. *Hamburg Wechsler Intelligenztest für Kinder –IV* (*HAWIK-IV*). (Huber, 2007).
62. von Aster, M., Neubauer, A. & Horn, R. Wechsler-Intelligenztest für Erwachsene: Übersetzung und Adaption der WAIS-III von David Wechsler [Wechsler intelligence test for adults: Translation and adaptation of WAIS-III by David Wechsler]. (Harcourt Test Services, 2006).
63. Petersen, A. C., Crockett, L., Richards, M. & Boxer, A. A self-report measure of pubertal status: Reliability, validity, and initial norms. *J. Youth Adolesc.* **17**, 117–133, doi:10.1007/BF01537962 (1988).

Acknowledgements

This work was supported by the German Ministry of Education and Research (BMBF grants # 01 EV 0711, 01 EE 1406B), the Deutsche Forschungsgemeinschaft (SFB 940/1), and the MedDrive Start Grant of the Medical Faculty of the Technische Universität Dresden. Eva Mennigen was supported by the Max Kade Foundation, NY, USA. We would like to thank our adolescent participants and their families.

Author Contributions

S.J., S.R., E.M., N.C.V. assessed and analyzed fMRI data. N.C.V. and J.S. analyzed behavioral and fMRI reliability data. N.C.V. wrote the main manuscript text and the methods section. M.N.S. conceived the study design. All authors provided comments on initial versions of the article.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-02334-7

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.