Developmental Cognitive Neuroscience 33 (2018) 17–26

Contents lists available at ScienceDirect

Developmental Cognitive Neuroscience

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

Test-retest reliability of longitudinal task-based fMRI: Implications for developmental studies

Megan M. Herting[a,*], Prapti Gautam[b,c], Zhanghua Chen[a], Adam Mezher[d], Nora C. Vetter[e,f,g]

[a] Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90032, United States
[b] Department of Psychology, University of Southern California, Los Angeles, CA 90089, United States
[c] Centre for Research on Ageing, Health, and Wellbeing, The Australian National University, Canberra, ACT, Australia
[d] Neuroscience Graduate Program, University of Southern California, Los Angeles CA 90007, United States
[e] Neuroimaging Center & Department of Psychiatry and Psychotherapy, Technische Universität Dresden, Germany
[f] Department of Child and Adolescent Psychiatry, Faculty of Medicine of the Technische Universität Dresden, Germany
[g] Department of Psychology, Bergische Universität Wuppertal, Germany

ARTICLE INFO

Keywords:
fMRI
Test-retest reliability
Intraclass correlation
Development

ABSTRACT

Great advances have been made in functional Magnetic Resonance Imaging (fMRI) studies, including the use of longitudinal design to more accurately identify changes in brain development across childhood and adolescence. While longitudinal fMRI studies are necessary for our understanding of typical and atypical patterns of brain development, the variability observed in fMRI blood-oxygen-level dependent (BOLD) signal and its *test-retest reliability* in developing populations remain a concern. Here we review the current state of test-retest reliability for child and adolescent fMRI studies (ages 5–18 years) as indexed by intraclass correlation coefficients (ICC). In addition to highlighting ways to improve fMRI test-retest reliability in developmental cognitive neuroscience research, we hope to open a platform for dialogue regarding longitudinal fMRI study designs, analyses, and reporting of results.

1. Introduction

The overarching question the field of developmental cognitive neuroscience attempts to answer is "What factors shape the development of our brain and behavior?" Functional Magnetic Resonance Imaging (fMRI) and neuropsychological research have provided a wealth of knowledge about the similarities and differences between child and adolescent brains compared to adult brains and their behavioral phenotypes (Casey et al., 2008; Blakemore, 2012; Crone and Dahl, 2012; Crone and Elzinga, 2015). However, these brain-behavior relationships have largely been studied using cross-sectional designs, which may not accurately describe **true developmental change**[1] within individuals. To address this concern, the field of developmental cognitive neuroscience is moving towards implementing longitudinal study designs, which can better capture within-subjects differences across child and adolescent development.

The growing awareness of the advantages offered by longitudinal design is evident by an increase in the number of longitudinal fMRI studies published in children and adolescents (ages 5–18 years), with most of these papers published within the past 5 years. Moreover, additional studies are expected given the number of large-scale consortium projects that have been funded globally in recent years. These include population-based longitudinal developmental MRI studies such as the IMAGEN Project (PL037286) by the European Commission's 6th Framework Program (IMAGEN, 2007) (N = 2000; 14 year olds over 5 years) that began in 2007, the National Consortium on Alcohol & Neurodevelopment in Adolescence (NCANDA, 2014) (NCANDA; N = 800 high-risk youth; 12–21 year olds over 3 years), as well as the recently established Adolescent Brain Cognitive Development (ABCD) (ABCD, 2015) initiative supported by the National Institutes of Health in the United States (N = 10,000; 9–10 year olds over 10 years).

As longitudinal designs become more widely used in developmental fMRI studies, another consideration for the field is the need for a better understanding of the variability observed longitudinally in fMRI signals and its test-retest reliability in developing populations. Test-retest reliability is the consistency of an assessment tool to produce stable

Abbreviations: ICC, intraclass correlation; HLM, hierarchical linear modeling

* Corresponding author at: 2001 N. Soto St, University of Southern California, Los Angeles, CA 90032, United States.
E-mail addresses: herting@usc.edu (M.M. Herting), prapss@gmail.com (P. Gautam), zhanghuc@usc.edu (Z. Chen), mezher@usc.edu (A. Mezher), nora.vetter@tu-dresden.de (N.C. Vetter).

[1] True developmental change – Throughout the paper we refer to the actual or real developmental process that occurs with age as "true developmental change". Regardless of our ability to measure it, we assume that there are true developmental changes that occur in brain activity (as measured by task-related fMRI BOLD signal) and behavior.

http://dx.doi.org/10.1016/j.dcn.2017.07.001
Received 23 March 2016; Received in revised form 29 June 2017; Accepted 5 July 2017
Available online 13 July 2017
1878-9293/ © 2017 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

results with each use (Khoo et al., 2006). That is, if an assessment tool is highly reliable, it will yield very similar results each time it is used, assuming there are no confounding factors during the interval of time between subsequent measurements.

In this article, we start with a brief overview of how longitudinal fMRI design and mixed effects modeling allow for more innovative approaches to studying individual differences in brain function across development. Next, we discuss the importance of test-retest reliability for task-based fMRI and summarize longitudinal studies that have examined task-based reliability, with a focus on how results from studies of children and adolescents with poor reliability metrics may be especially difficult to interpret. We conclude by reviewing existing approaches used to minimize factors that may contribute to poor fMRI test-retest reliability in developmental cognitive neuroscience research.

2. Measuring developmental change in task-based fMRI BOLD using mixed effects models

As detailed by a recent timely review (Crone and Elzinga, 2015), cross-sectional studies have several limitations. For example, they are likely to suffer from cohort effects and are unable to assess causal factors. Longitudinal studies, however, allow for determining how much children and adolescents differ from one another (between-subject variance), but also how much a particular child or adolescent changes over time (within-subject variance) (Singer and Willett, 2003). Below we present hypothetical data to exemplify longitudinal fMRI BOLD signal in each subject as a function of (measurable) intercepts and slopes using mixed effects models.

In regard to distinguishing within-subject change and between-subject differences in task-related fMRI BOLD signal (either within a voxel or a region of interest (ROI)), mixed effects modeling (Fig. 1) is a powerful statistical method for studying how children and adolescents differ in how they change over time. Mixed effects modeling expands upon multiple regression for repeated-measures and uses random effects to differentiate between- and within-subject variance (Singer and Willett, 2003). That is, this technique uses a random intercept and slope to model an initial state (i.e. intercept) and generate a linear growth trajectory of fMRI BOLD signal (i.e. slope; see Fig. 1) for each individual. It also evaluates how additional variables (i.e. sex, genotype, stress) may explain the intercept and/or slope of the fMRI BOLD signal over time (Singer and Willett, 2003). In particular, the mixed effects model can be expressed as two-level models such that the full mixed effect model:

$$Y_{ij} = a_0 + \boldsymbol{\beta' Z_{ij}} + \boldsymbol{a_1 X_i} + b_0 t_{ij} + \boldsymbol{b_1 X_i} \times t_{ij} + b_i t_{ij} + a_i + \varepsilon_{ij} \quad (1)$$

can be broken down into:

Level 1: $Y_{ij} = \boldsymbol{\alpha_i} + \boldsymbol{\beta' Z_{ij}} + \boldsymbol{\beta_i} t_{ij} + \varepsilon_{ij}$

Level 2 (a): $\boldsymbol{\alpha_i} = a_0 + \boldsymbol{a_1 X_i} + a_i$

Level 2 (b): $\boldsymbol{\beta_i} = b_0 + \boldsymbol{b_1 X_i} + b_i$

where i indicates individual i = 1,2,…,n; t_{ij} represents follow-up time at visit j = 0,1,2, ….; a_i and b_i are random variations of the intercept and the slope of time, which are assumed to follow bivariate normal distributions of $[a_i, b_i] \sim N(0, \Sigma_{a,b})$; ε_{ij} is random error. X is an additional matrix of time-independent covariates and can include multiple factors from $X(1)$, $X(2)$, …, $X(p)$ (e.g. sex, genotype). Z_{ij} is a matrix of time-dependent covariates and also include multiple factors (e.g. stress or sex hormone levels). In other words, the Level 1 model refers to the within-subject change model and describes whether time-dependent covariates $\boldsymbol{Z}$ are associated with the developmental changes (e.g. the variation within the individual over time) for each individual. The Level 2 model captures whether variations of the initial state (Level 2(a)) and the rates of change (Level 2(b)) across individuals are associated with variables of interest (e.g., $\boldsymbol{X}$). With ≤3 timepoints per subject, only a linear Level 1 model can be fit to a single subject's data, whereas with more than 3

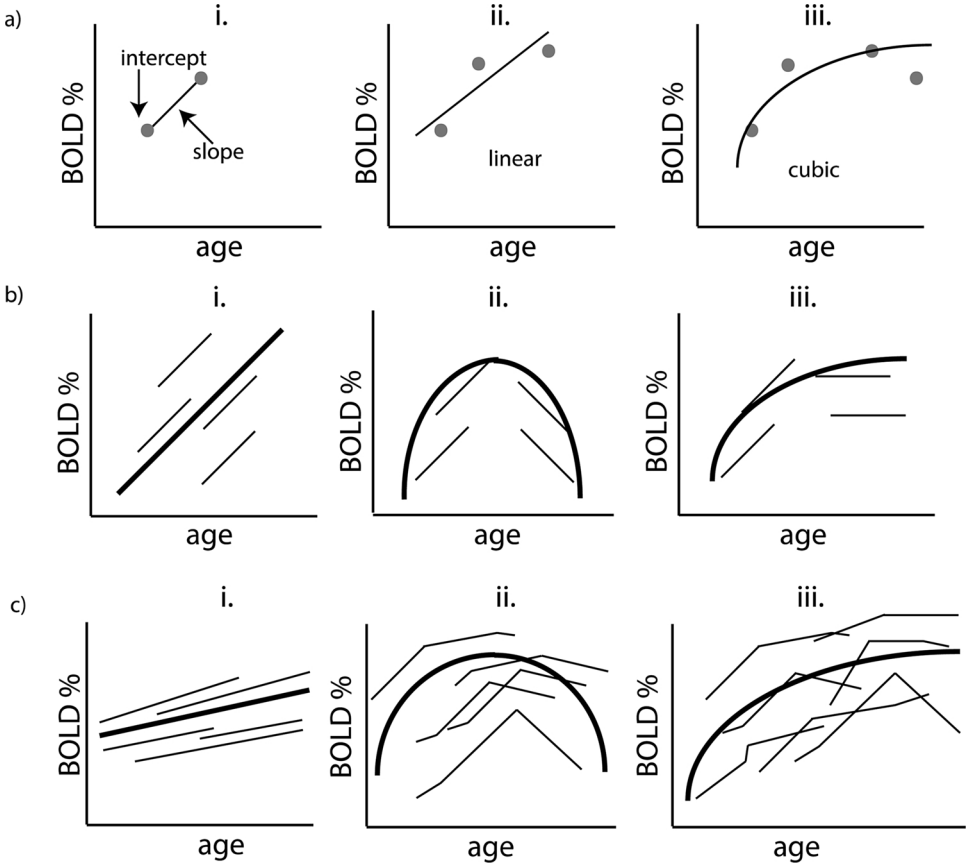

Fig. 1. Longitudinal trajectories measured at the individual level and at the group level using mixed effects modeling.
a. Level 1 Model: When the same child or adolescent is measured over time (e.g. age), individual parameters can be estimated that characterize the starting point (intercept) and the change (slope) of the fMRI BOLD signal for that specific subject (a.i.). Individual estimates of the longitudinal pattern of the fMRI BOLD signal will have greater precision when the number of timepoints is increased. With three timepoints, a simple linear model can be fit to estimate linear trajectory (a.ii), whereas with 4+ timepoints non-linear growth trajectories of a child or adolescent (such as quadratic and cubic) can be estimated (a.iii.).
b. Level 2 Model: Linear (b.i.), quadratic (b.ii.) and cubic (b.iii.) between-subject estimates can be assessed if a large age-range of children and adolescents are included in the sample. In this scenario, including a wide age-range allows more complex models (i.e. quadratic, cubic, etc.) to be fitted at a between-subject level.
c. i/ii/iii. With the inclusion of more timepoints, both linear (c.i) and non-linear (c.ii/iii) models can be fitted for both within- and between-subject levels.

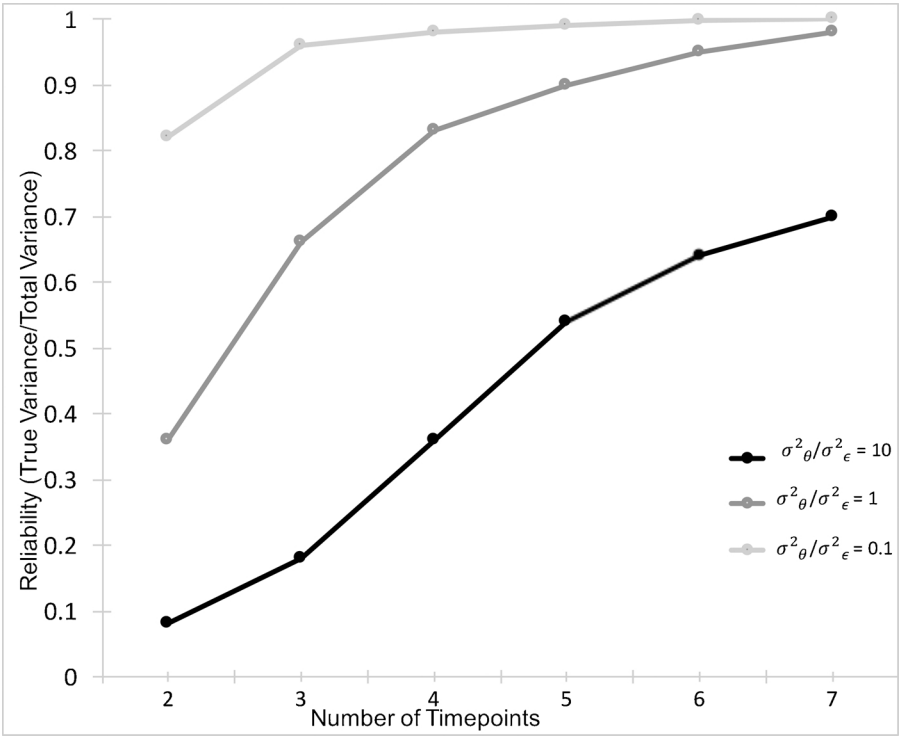

Fig. 2. The reliability of the change estimate (i.e. slope) increases with the number of timepoints collected per subject. σ^2_θ represents the between-subjects differences in true change (i.e. slope) and σ^2_ε represents the measurement error variance. Adapted from Willett (1989).

timepoints, more complex shape trajectories (e.g. quadratic, cubic) can be used to examine within-subject changes over time (Fig. 1).

Using this model, both the intercept and/or slope can be either fixed or varied across individuals. A random intercept and fixed slope model allows for the intercept to vary (i.e. each child and adolescent differs in BOLD signal at baseline), but assumes that the slopes are fixed (i.e. each subject's BOLD signal trajectories are similar over time). A fixed intercept and random slope model allows for the slopes to vary (i.e. each subject's BOLD signal has different trajectories over time), but assumes that the intercept is fixed (i.e. each child has the same fMRI BOLD signal at baseline). A random intercept and random slope model provides the most flexibility, allowing both the intercept and the slope to vary (i.e. each child or adolescent has different fMRI BOLD signal at baseline and different longitudinal trajectories of BOLD signal over time).

To accurately and reliably estimate *change* (i.e. Level 2b; the slope parameter β_i) using mixed effect models, it is critical to minimize the measurement error term (ε). If the random measurement error variance is large compared to the inter-individual variation in the true change score, then the ability to reliably estimate the slope will be poor (Fig. 2). Alternatively, when the inter-individual variance in the true score is large compared to the measurement error, the change parameter estimate will be more reliable. Collecting more timepoints per person can significantly improve the estimates of change in the Level 1 model (Singer and Willett, 2003). That is, the parameter estimates improve with the collection of additional timepoints (Fig. 2). In fact, for any level of measurement error (i.e. high, medium, or low), increasing from 2 to 3 timepoints leads to a more reliable estimate of change β_i.

To adapt this concept to developmental fMRI studies, consider the following example: the goal of a study is to examine how a child or adolescent's amygdala activity (e.g. BOLD signal) changes in response to emotional faces over time, and to determine how between-subject factors like biological sex (time-independent) and stress (time-dependent) influence longitudinal changes in neuronal activity. To accomplish this, the study utilizes an accelerated longitudinal design, which enrolls children at different ages at baseline and follows them every 2 years. Using mixed effect modeling, we may then answer the following set of questions:

1. What is the fMRI BOLD signal in the amygdala for a child at the beginning of the study (e.g. baseline) and does the BOLD signal of the amygdala change in that child over time? (Intercept and slope (Level 1))
2. Is the fMRI BOLD signal different between boys and girls at baseline of the study (i.e. do the intercepts differ between subjects)? (Estimated by the parameter in Level 2 (a))
3. Does the fMRI BOLD signal change differently between boys and girls over time (i.e. does the mean slope of time differ between boys and girls)? (Estimated by the parameter in Level 2 (b))
4. Is the change in fMRI BOLD signal over time associated with the change in stress levels over time? (Estimated by the parameter in Level 1)
5. How does the baseline fMRI BOLD signal in the amygdala directly affect the longitudinal trajectory of the BOLD signal in children? (Treat baseline value of the fMRI BOLD signal as one of the time-independent factors in $\boldsymbol{X}$)

In this example, mixed effects modeling is a much more appropriate method than earlier and more commonly used methods like univariate repeated measures analysis of variance (ANOVA) and multivariate repeated measures analysis of variance (MANOVA). This is because (i) it can incorporate restricted maximum likelihood (REML) to handle incomplete (i.e. missing) data as well as irregularly spaced timepoints; (ii) it can estimate rate of change (slope) for each individual instead of average group slope, and (iii) it can assess how continuous and time-varying covariates, such as stress in this example, influence longitudinal relationships (Singer and Willett, 2003).

3. Test-retest reliability of fMRI BOLD signal

As measured, the observed fMRI BOLD signal is comprised of both the true BOLD signal value plus error. Continuing with our previous example looking at changes in amygdala activity in response to emotional faces, let us assume that the fMRI BOLD signal in the amygdala for this task does in fact change within-subjects over time (e.g. with age) across childhood and adolescence. One can then characterize the

Table 1
Study design recommendations to minimize sources of variation in fMRI studies.

Source	Description	Recommendation
MR related		
Scanner	Machine characteristics and performance (i.e. changes in scanner or changes in software or hardware on the same scanner), scanner stability	Perform data quality measurements (e.g., signal to noise) with phantoms before each data collection
Acquisition method	Pulse sequence and imaging parameters	Investigators should schedule regular maintenance
Placement	Differential subject position in bore	Longitudinal studies and or different sites agree upon "range of values" that all scanners should adhere in order to standardize measurements
Subject related		
Subject	Individual differences in physiology, responses and hormonal rhythms	Conduct fMRI at similar times to minimize fluctuations due to circadian rhythms
Sample	Cohort size and composition	Determine by task design/cognitive construct/sample characteristics
Intrinsic	Noise and other unaccounted variation	Noise due to intra-individual variability could be minimized by increasing measurement occasions
Task related		
Longitudinal processing	Voxel registration across timepoints to one another and to same anatomical location	Standardized MRI acquisitions followed by standardized pre-processing of imaging data to minimize differences
Motion during scans	Differences in motion across timepoints	Surface-based registration/analysis may help to reduce the influence of changes in cortical thickness which occur with development
Practice effects	Subjects might get better at the task at subsequent visits	Have an alternate version of the task or use adaptive methods where task difficulty is matched to each subject
Block between session	Variation across responses to each task presentation (attention, arousal, caffeine, etc.); non-task related cognitive processes; changes in cognitive strategy over time; task comprehension, attention and arousal	Task comprehension is easily solved by practice sessions before going into the scanner
Task	Block vs. event-related designs; target region	Event-related designs generally need longer task designs; investigator should consider the best approach
Time between session/ Lag time	Temporal artifacts (e.g. drift, low-frequency oscillations, etc.)	Times between sessions should be small enough that there is no "developmental change". This would vary depending on the cognitive construct and would need to be trialed by the investigator

change in amygdala BOLD signal as *"true developmental change"*. To accurately measure this true individual developmental change, one must distinguish true developmental change from other sources of variability. Hence, our goal is to identify and adjust for variability from other sources that are not considered true developmental changes in the analyses. If the fMRI BOLD signal for a given task (e.g. emotional stimuli) in a specific ROI (e.g. amygdala) has poor test-retest reliability, this will affect the measurement error term during statistical testing (Level 1 model error term: ε_{ij}). If one does not consider test-retest reliability of the BOLD signal in developmental fMRI longitudinal studies, one risks conflating poor measurement reliability in attempts to characterize "true developmental change" (McArdle and Woodcock, 1997).

3.1. Intraclass correlation coefficient (ICC)

Test-retest reliability, or the consistency of the fMRI BOLD signal over time, can be measured quantitatively by computing an intraclass correlation coefficient (ICC) (Bennett and Miller, 2010). The ICC equation is listed below in Eq. (2) (Shrout and Fleiss, 1979).

$$ICC = \frac{between\ \ subjects\ \sigma^2}{(between\ \ subjects\ \ \sigma^2 + pooled\ \ within\ \ subjects\ \ \sigma^2)} \quad (2)$$

ICC assesses the similarity between two measurements; in this case it can be used to estimate the consistency of two measurements separated by time. Thus, ICC takes into account both the within-subject and between-subject variances in order to provide a ratio of the variance between subjects to the total variance. ICC values range between 0 and 1 and are commonly classified as poor (< 0.4), fair (0.41–0.59), good (0.6–0.74), and excellent (0.75–1) (Cicchetti and Sparrow, 1981; Cicchetti, 2001). Higher ICC values reflect greater test-retest reliability, or a more stable measurement between two timepoints. For example, an ICC of 0.82 can be interpreted as 82 percent of the variance being due to "true" variance between individuals, whereas the other 18 percent of variance is due to measurement error and/or within-subject variability (Bartlett and Frost, 2008).

It is important to note that ICC is distinctly different from a traditional Pearson's correlation. One key difference between these two statistics is that for ICC the data is centered and scaled using a pooled mean and standard deviation, whereas Pearson's correlation centers and scales each variable by its own mean and standard deviation. ICC also provides a more accurate estimate as it can differentiate both systematic variation and average consistency over time (Hunt, 1986). As the ICC is strongly influenced by trait variance of sampled data, ICC measured for different populations might not be comparable. This is because between-subject and within-subject variance might be different for different sub-populations. For example, the sample might be different (between-subject variance) and/or the people might be changing differently within the sample (within-subject variance). As such, ICC should be estimated for each population separately, unless it can be verified that the between-subject and within-subject population variances are similar to previously published results.

3.2. Test-retest reliability of task-based fMRI in adults

Using ICCs, recent efforts have examined test-retest reliability of task-based fMRI BOLD signal in adults. Bennett and Miller performed a meta-analysis of 13 fMRI studies between 2001 and 2009 that reported ICCs (Bennett and Miller, 2010). ICC values ranged from 0.16 to 0.88, with the average reliability being 0.50 across all studies. Others have also suggested a minimal acceptable threshold of task-based fMRI ICC values of 0.4–0.5 to be considered reliable (Aron et al., 2006; Eaton et al., 2008). However, most of these studies consisted of small samples (N = 10–30) of young adults (e.g. Bennett and Miller, 2010; Brandt et al., 2013; Lipp et al., 2014). Moreover, Bennett and Miller, as well as a more recent review (Dubois and Adolphs, 2016), highlight that reliability can change on a study-by-study basis depending on several methodical considerations.

Besides general study design, Genovese et al. (1997) have outlined a number of methodical sources that can influence test-retest reliability in fMRI, which we have summarized and expanded upon in Table 1. When designing an fMRI study, a number of these sources of variation can be standardized both between-subjects and within-subjects to improve test-retest reliability of results. Reduction in MRI scanner related variance, optimized acquisition parameters, as well as well-designed fMRI

task paradigms all help to improve the signal-to-noise ratio (SNR) in the BOLD signal. In turn, the better our ability to increase the functional signal amplitude and decrease noise across timepoints, the better our estimation of the signal and its reliability over time (Bennett and Miller, 2010). Temporal and spatial noise in fMRI often results from intrinsic thermal noise from the scanner and subject, system noise due to scanner hardware, artifacts from physiological processes of the subject, and variability of neural activity associated with non-task related neural activity. Thus, optimizing scanner sequences, reducing scanner artifacts, and ensuring system stability with phantoms should lead to better SNR (Huettel et al., 2004) and, subsequently, increase test-retest reliability.

Beyond functional SNR, the test-retest reliability of BOLD signal is also dependent on the fMRI paradigm. For example, in adult studies with only 2–3 days between timepoints, large differences in reliability have been reported for various fMRI tasks (i.e. motor, language, etc.) using the same sample of subjects (Gorgolewski et al., 2013a, 2013b). ICC estimates also vary by fMRI task design, with greater reliability reported for block versus event-related fMRI paradigms (Bennett and Miller, 2013). Moreover, reliability metrics are likely to vary by cortical regions for a given fMRI task. In a recent reliability analysis of a verbal working memory task in adults (8 participants, 2 scans per site, and 8 site locations), the proportion of variance explained by the person, the day, and the scanner site (and possible interactions) widely varied across 10 cortical ROIs (Forsyth et al., 2014). Lastly, truly understanding test-retest reliability for task-related fMRI likely requires examining ICC values for each fMRI task condition as well as the fMRI contrast of interest. For example, task-based fMRI analyses examine the difference in BOLD signal between two (or more) task conditions, including a task of interest (Task A) and a control task (Task B), in order to subtract neuronal activity that is common between the two tasks and highlight neuronal activity that is task-specific. Thus, the importance of choosing an appropriate control task, and its consequences on the results and interpretation, has been previously acknowledged for functional imaging (Church et al., 2010). Similarly, there is evidence that a contrast (one condition versus another) versus implicit baseline might lead to different reliability estimates (Bennett and Miller, 2013).

Taken together, test-retest reliability of the fMRI signal is contingent upon optimization and standardization of the scan protocol, the fMRI task paradigm, and the ROIs. It is important to note, that ICC estimates for adult studies are based on short intervals between scans (one to several weeks)(e.g. Bennett and Miller 2010; Brandt et al., 2013; Lipp et al., 2014), and that fMRI measurements that are taken closer in time are more likely to be similar. This is in stark contrast to task-based fMRI studies of children and adolescents, that often examine and report ICC values for test-retest reliability as part of a longitudinal study with a substantial delay between measurements. Thus, adult task-based fMRI ICC values may not generalize to longer time intervals and − more importantly − the reliability in child and adolescent samples may not be comparable to the reliability thresholds seen in adult samples. Furthermore, it is possible that BOLD signal reliability can differ substantially between specific developmental stages (e.g. children, adolescents, adults (e.g. Koolschijn et al., 2011)).

3.3. Test-retest reliability of task-based fMRI in children and adolescents

While a number of longitudinal task-based fMRI studies on children and adolescents have been published, few studies have reported test-retest reliability metrics for the task-based fMRI BOLD signal associated with the fMRI task paradigm. To our knowledge, ICC values have been reported for within-subject variance (Level 1 model) for 12 studies (Koolschijn et al., 2011, Britton et al., 2013; Ordaz et al., 2013; van den Bulk et al., 2013; Braams et al., 2015; Paulsen et al., 2015; Qu et al., 2015; Vetter et al., 2015; Peters et al., 2016; White et al., 2016; Vetter et al., 2017). Details on the respective samples, fMRI task paradigm, ROIs and contrasts of these studies can be found in Table 2. Similar to adult studies, ICC values calculated over a pre-defined ROI have been shown to fall largely within the fair range, although with notable differences across fMRI task and brain region (e.g. Fig. 3). Reliability tends to be best (good to excellent) for occipital regions (Koolschijn et al., 2011; Vetter et al., 2015; Vetter et al., 2017) and fair to poor for subcortical regions, including the amygdala, nucleus accumbens, and putamen (Ordaz et al., 2013; van den Bulk et al., 2013; Braams et al., 2015; Qu et al., 2015; Vetter et al., 2015; White et al., 2016; Vetter et al., 2017). However, it should be noted that when a voxelwise approach was performed (as opposed to an anatomically defined ROI), higher reliability estimates (excellent) were reported for BOLD signal in portions of the amygdala and para-hippocampus when using an emotional dot-probe task (Britton et al., 2013).

Similar to adult studies, the reliability of the BOLD signal is task and contrast of interest specific in developing populations as well. That is, ICC values are different in the same ROI depending on the fMRI task and the contrasts of interest. In a recent study, Vetter et al. (2017) investigated reliability for three tasks (emotional attention task, cognitive control task, reward task) in the same sample of N = 104 adolescents at age 14 and again at age 16, with reliability estimates ranging from poor to excellent depending on the ROI and task. Whole-brain ICC estimates were larger for the inter-temporal choice paradigm (reward task), followed by the emotional attention task, and then the cognitive control task. Moreover, across all three fMRI tasks, good to excellent reliability was found for the superior occipital cortex, whereas ICC values for the other ROIs were variable by task condition.

One important point to note is that ICC reliability estimates from these longitudinal studies of development, which tend to have substantial delays between timepoints, are harder to interpret compared to adult fMRI BOLD reliability estimates. Similar to any outcome variable, the test-retest reliability of the fMRI BOLD signal can be influenced if the persons being studied change dramatically between the test and retest points (McArdle and Woodcock, 1997). In adult studies of fMRI reliability, the duration between scans is shorter and ICC values can more easily be interpreted as how reliable the BOLD signal is for a given task. In longitudinal studies of children and adolescents, however, maturation may influence the magnitude of the BOLD signal as well as the specific brain regions involved in a particular task (which one might expect as a function of development across childhood and adolescence); this maturation over time should lead to lower ICC scores. Thus, lower test-retest reliability of BOLD signal in longitudinal studies may reflect 1) poor consistency of the fMRI measurement (BOLD signal) itself and/or 2) that the subjects changed over time (which is what we hope for when studying age related development). To overcome this challenge of how to accurately interpret low reliability estimates, existing longitudinal fMRI studies have investigated the ICCs of different ROIs that are assumed either to continue to develop with age (e.g. subcortical and cortical areas: amygdala, prefrontal cortex) or to remain stable with age (e.g. occipital regions). For example, two studies have examined ICCs in various ROIs across groups of participants at different ages using task-based fMRI paradigms (Koolschijn et al., 2011; Peters et al., 2016). Koolschijn et al. (2011) compared reliability in 8–11 year-olds, 14–15 year-olds and 17–25 year-olds in a rule-switch task with a between scan interval of ~3.5 years. Reliability estimates were different for each cortical ROI by age group when examining the BOLD contrast between the first warning of a rule change vs. positive feedback. Poor reliability was observed in all cortical ROIs for the 8–11 year-olds, but only in the insula for the 14–15 year-olds and only in the orbitofrontal gyrus for the 17–25 year-olds. For the remaining cortical ROIs (parietal, precuneus, angular gyrus, and anterior cingulate cortex), fair to good reliability was observed for both the 14–15 and 17–25 year-olds. Peters et al. (2016) employed a feedback learning fMRI task in participants with approximately 2 years between scans. ICC values for the BOLD contrast between learning vs. application trials were found to be poor in the superior parietal cortex for the 8–12 year-olds and in the dorsolateral prefrontal cortex for the 17–25 year-olds (Peters et al., 2016).

Table 2
ICC values reported in developmental task-based longitudinal fMRI studies. Reliability: poor (< 0.4), fair (0.41–0.59), good (0.6–0.74), and excellent (0.75–1) (Cicchetti, 2001). Region approach: ICC calculated for 1) ROIs − anatomically derived ROIs, 2) fMRI ROIs − functionally derived ROIs, or 3) at a voxelwise level. Numbering within the ICCs/ROIs column highlights differences between age groups or fMRI contrasts for a given study.

Author	Sample	Task	Region Approach/Contrast	ICCs / ROIs
Koolschijn et al. (2011)	N = 10; 8–11 yrs N = 12; 14–15 yrs N = 10; 18–25 yrs Design: 2 waves ~3.5 yr interval	Cognitive switch task (rule change)	ROIs Contrast: First warning > positive feedback	1) 8–11 yrs: Poor: all ROIs 2) 14–15 yrs: Poor: insula Fair: frontal gyrus Good: parietal cortex, precuneus, angular gyrus, anterior cingulate cortex 3) 18–25 yrs: Poor: orbitofrontal gyrus Fair: insula, parietal cortex, frontal gyrus, angular gyrus, anterior cingulate cortex Good: parietal, precuneus, frontal gyrus
van den Bulk et al. (2013)	N = 27; 12–19 yrs Design: 3 waves ~6 month interval	Face attention paradigm (fearful, happy, neutral)	ROIs Contrast: All faces > fixation	Poor: medial prefrontal cortex, R lateral prefrontal cortex, amygdala Fair: L lateral prefrontal cortex Excellent: inferior occipital cortex
Britton et al. (2013)	N = 12; 8–17 yrs Design: 2 waves 121 ± 50 day interval	Emotional dot-probe task (angry, fearful, neutral faces)	Voxelwise ICC > 0.56 Contrast: Masked or Unmasked Angry Bias Fearful Bias	1) Unmasked Angry Bias: Good: inferior frontal gyrus 2) Unmasked Fearful Bias: Excellent: inferior frontal gyrus 3) Masked Fearful Bias: Excellent: L amygdala/para-hippocampus 4) Fearful incongruent > fixation: Excellent: L/R amygdala/para-hippocampus
Ordaz et al. (2013)	N = 123; 9–26 yrs Design: Up to 6 waves ~1 yr interval	Visual anti-saccade task	ROIs Contrast: Antisaccade > fixation	Poor: supplementary and frontal eye-field, pre- supplementary motor area, posterior parietal cortex, dorsolateral and ventrolateral prefrontal cortex, dorsal anterior cingulate cortex, putamen Fair: R dorsolateral prefrontal cortex
Braams et al. (2015)	N = 238; 8–25 yrs Design: 2 waves ~2 yr interval	Reward task (gambling game)	ROIs Contrast: Win > lose	Poor: nucleus accumbens
Qu et al. (2015)	N = 23; 15–17 yrs Design: 2 waves ~1.5 years	Balloon analog risk task	fMRI ROIs Contrast: Cash out > baseline	Poor: ventral striatum, dorsolateral prefrontal cortex
Vetter et al. (2015)	N = 144; 14 yrs Design: 2 waves ~2 yr interval	Emotional attention (IAPS matching task, negative, positive, neutral)	fMRI ROIs Contrast: Negative attended pictures > baseline	Poor: medial prefrontal cortex, inferior frontal gyrus, anterior cingulate cortex, amygdala Excellent: fusiform gyrus
Braams and Crone (2016)	N = 254; 8–27 yrs Design: 2 waves ~2 yr interval	Heads/tails gambling task	ROIs Contrast: Friend > self	Poor: ventral medial prefrontal cortex, precuneus, temporoparietal junction
McCormick et al. (2016)	N = 20; 14 yrs Design: 2 waves ~1 yr interval	Go/nogo task	fMRI ROI Contrast: Successful nogo > baseline	Poor: ventral lateral prefrontal cortex
Peters et al. (2016)	N = 74; 8–12 yrs N = 89; 13–16 yrs N = 45; 17–25 yrs Design: 2 waves ~2 yr interval	Feedback learning task	ROIs Contrast: Learning > application	1) 8–12 yrs: Poor: superior parietal cortex Fair: dorsolateral prefrontal cortex, supplementary motor area, anterior cingulate cortex 2) 13–16 yrs: Fair: dorsolateral prefrontal cortex Good: superior parietal cortex, supplementary motor area, anterior cingulate cortex 3) 17–25 yrs: Poor: dorsolateral prefrontal cortex Fair: superior parietal cortex, supplementary motor area, anterior cingulate cortex
White et al. (2016)	N = 39; 10–17 yrs Design: 2 waves 9.36 ± 2.09 week interval	Emotional dot-probe task (angry and neutral)	Voxelwise ICC > 0.41 ROI of amygdala Contrast: All conditions > baseline Angry Bias	1) All conditions > baseline: Good: inferior, precentral and middle frontal gyrus Excellent: middle frontal gyrus 2) Angry Bias: Poor: amygdala

(continued on next page)

Table 2 *(continued)*

Author	Sample	Task	Region Approach/Contrast	ICCs / ROIs
			Angry > neutral	Fair: L claustrum, L insula, L inferior frontal gyrus 3) Angry > Neutral: Poor: amygdala
Vetter et al. (2017)	N = 104, 14 yrs Design: 2 waves ~2 yr interval	Cognitive control (interference switching task) Reward (intertemporal choice) Emotional attention (IAPS matching task, negative, positive, neutral)	fMRI ROIs Contrast: Switch incongruent > baseline Contrast: Intertemporal decision phase > baseline Contrast: Negative attended pictures > baseline	Poor: R dorsolateral prefrontal cortex, R dorsal anterior cingulate cortex Fair: posterior parietal cortex, L dorsolateral prefrontal cortex, L dorsal anterior cingulate cortex Good: superior occipital cortex Poor: L ventral striatum Fair: anterior cingulate cortex, R ventral striatum Excellent: L/R superior parietal lobe, L/R fusiform gyrus, superior occipital cortex Poor: medial prefrontal cortex, inferior frontal gyrus, anterior cingulate cortex, amygdala Good: R superior occipital cortex Excellent: fusiform gyrus, L superior occipital cortex

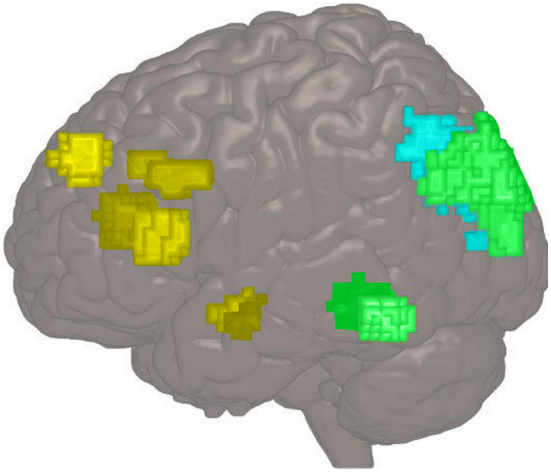

Fig. 3. Test-retest reliability for different cortical ROIs in an emotional attention task. Image from adolescents scanned at age 14 and again at age 16 in a study by Vetter et al. (2015). All regions are depicted on the rendered surface bilaterally (to show both sides). Yellow represents areas with poor reliability (ICCs: < 0.4) (medial prefrontal cortex, anterior cingulate cortex, bilateral inferior frontal gyrus and bilateral amygdalae); blue represents areas with good reliability (ICCs: 0.6-0.74) (right superior occipital cortex); green represents areas with excellent reliability (ICCs: 0.75-1) (bilateral fusiform gyrus, left superior occipital cortex).

Alternatively, 13–16 year-olds had good reliability estimates for ROIs in the superior parietal cortex, supplementary motor area, and anterior cingulate cortex, whereas reliability was fair in these regions for the 8–12 and 17–25 year-olds. The differences in ICC between brain regions and/or age groups have thus been taken to represent greater developmental change. For example, in Koolschijn et al. (2011), the lower reliability values seen in younger children was interpreted as possibly reflecting larger maturation processes over the 3.5 year interval than those individuals who began the study at 14–15 or 17–25 years old (since these groups were found to have higher reliability estimates over the 3.5 year interval). This approach of comparing ICC values assumes that brain areas that are “developmentally sensitive” (e.g. prefrontal cortex or amygdala) have lower ICC values, whereas “developmentally insensitive” brain regions (e.g. occipital lobe) have higher ICC values when fMRI BOLD signal is repeatedly measured across a large developmental window.

While interesting, interpreting low reliability estimates of the BOLD signal from longitudinal designs with long delays between measurements as a ‘proxy for development’ (Koolschijn et al., 2011), may be misleading. A number of other factors may lead to lower ICC values (see Section 4 below). The ICC of the BOLD signal at shorter delays is first needed to establish reliability for any given task and/or ROI, in order to accurately interpret a low ICC value as a proxy for development. However, to our knowledge, only three studies to date have examined task-based fMRI reliability at relatively shorter delays between measurements in children and adolescents. Van den Bulk (2013) assessed adolescents (ages 12–19 years) within three to six months of their first visit using an emotional face paradigm. Poor to fair reliability was seen in the amygdala and prefrontal cortex ROIs, but excellent reliability for an occipital control ROI. Two additional studies using the emotional dot probe task reported poor amygdala reliability in 39 adolescents with an approximate 9 week scan interval (White et al., 2016), but good to excellent reliability in portions of the amygdala at a slightly longer delay of 121 days in 12 adolescents (Britton et al., 2013). While it is feasible that maturation may happen on a shorter time scale, these mixed findings of test-retest reliability in children and adolescents suggest that the BOLD signal may be less reliable in general for these tasks and/or ROIs. If this is the case, these findings are in line with those of Plichta et al., who also reported low reliability in the amygdala for an emotional task in adults (Plichta et al., 2014). Overall, test-retest reliability in task-based fMRI studies in children and adolescents are still scarce. Moreover, the meaning of low reliability in longitudinal fMRI studies with longer scan intervals should be interpreted cautiously given that reliability estimates of the BOLD signal itself and true developmental change are inherently confounded in these studies and difficult to tease apart. Poor test-retest reliability of the BOLD signal in longitudinal studies may reflect poor consistency of the fMRI measurement (BOLD signal) itself, represent true developmental change, or some combination of both.

4. Improving our understanding of test-retest reliability of task-based fMRI in children and adolescents

Although they have been seldom reported in the literature, reliability estimates of task-based fMRI are very important in deciphering meaningful developmental changes. For the reasons highlighted above, future task-based longitudinal test-retest fMRI studies are warranted in children and adolescents to more fully characterize the reliability of the BOLD signal. In this section, we discuss factors that should be carefully considered to improve reliability and to ensure more reproducible and generalizable findings in developmental fMRI experiments.

4.1. Length of time between measurements

Without knowing if an fMRI paradigm will show similar test-retest estimates across a range of retest intervals, it is difficult to decipher the meaning of ICC reliability values based on the developmental longitudinal fMRI studies to date. When establishing the reliability of an

fMRI task paradigm, an assumption is that the underlying fMRI construct for the paradigm does not change. Thus, in order for this assumption to be true in developmental populations, the length of time between test-retest reliability measurements has to be short. As mentioned, only 3 studies to our knowledge have examined reliability in a developmental sample within a short time frame (3–6 months) (Britton et al., 2013; van den Bulk et al., 2013; White et al., 2016). Despite this short interval between measurements, these studies reported variable reliability estimates, including some very poor ICC values for both cortical and subcortical ROIs (van den Bulk et al., 2013; White et al., 2016). These low reliability estimates for two closely measured timepoints could be due to a number of reasons (i.e. see Table 1 and see below). Alternatively, although perhaps less likely, it could be that the development of emotional-related attention processes may occur in as little as three to six months. This latter idea highlights the fact that the optimal time frame to capture test-retest reliability for task-based fMRI paradigms in children and adolescents is currently unknown. Given the dynamic neurodevelopmental trajectories of brain regions and cognitive abilities across childhood and adolescence, the duration between timepoints needed to reduce a developmental confound on reliability measurements is likely to vary depending on sample age and the cognitive or behavioral function being assessed by the task. That is, effects of true developmental change may be minimal in easy tasks (i.e. finger tapping) or baseline (i.e. fixation or rest) conditions compared to tasks that require higher cognitive and emotional processes that may continue to develop throughout childhood and adolescence. Even at longer delays, one might expect ICC values may be higher for control or baseline conditions as compared to other fMRI contrasts of interest. A short time interval of 1 day, 1 week, or 1 month would seem reasonable to minimize the confounds of "development", but empirical reliability data from developmental fMRI studies are necessary to determine the duration between timepoints needed for different fMRI task paradigms to assess reliability without also including developmental change in the outcome variable.

4.2. Practice effects, subject compliance, and cognitive strategies

Both practice effects and changes in cognitive processes to complete a specific task can directly impact reliability measurements in longitudinal child and adolescent studies. Practice effects can invariably occur in test-retest or longitudinal contexts as individuals over any age get the same tests over time. As a result, an individual's performance on a task, and subsequent brain activity, can be influenced by 1) familiarity with the actual test content, 2) familiarity with the MRI environment, or 3) improvements in cognitive or test-taking strategies. To reduce practice effects due to familiarity of test content, it is best to have alternative versions of the fMRI task paradigm that are counterbalanced across timepoints, such that the exact same test is not repeated and order effects are minimized (Beglinger et al., 2005). Dealing with the practice effects due to familiarity with the test environment, or, in other words, a subject's habituation to being in an MRI scanner, is more difficult to control. To reduce the effect of this confound, many studies perform mock scanning to familiarize all participants with the MRI environment and/or collect physical and/or psychological markers of stress and anxiety. If there is a long interval between measurements, developmental processes (e.g. based on age) are also likely to result in better subject compliance and improved cognitive strategies for many cognitive and emotional tasks (Schlaggar et al., 2002; Church et al., 2010). Thus, incorporating a control task into the study design could be useful, especially if it has the same level of difficulty as the task of interest but has little to no sensitivity to development. Moreover, direct assessment of cognitive development and strategy implementation could be assessed by examining changes in task performance as well as asking about strategies utilized during the paradigm. These measurements could then be used to see if they relate to reliability estimates within ROIs or task conditions (White et al. 2016).

4.3. Motion

Although the topic of motion is not specific to longitudinal task-based fMRI, it still requires consideration in the context of understanding true developmental change and establishing reliability estimates. Unsurprisingly, motion compliance is one of the more difficult challenges to optimizing MRI signal-to-noise in developmental studies (Church et al., 2010). It is well understood that motion has potentially devastating effects on reliability outcome measurements. For example, one study has quantified how motion impacts fMRI reliability measurements in adults (N = 10, age range: 50–58, reliability scans taken 2–3 days apart) (Gorgolewski et al., 2013a, 2013b). Using two metrics of motion (i.e. total displacement and stimulus by motion correlations), it was shown that correlations between motion and stimulus presentation had large effects on reliability estimates (20–23% of the explained variance), and significantly decreased test-retest reliability (Gorgolewski et al., 2013a, 2013b). However, the confound(s) of motion on task reliability may be especially problematic for longitudinal developmental fMRI studies (Power et al., 2012). As children and adolescents mature they often show less motion during MRI acquisitions (Blumenthal et al., 2002). With longer intervals between measurements, the subject may move significantly less at time 2 than at time 1, yielding higher SNR at time 2. Thus, with long intervals between fMRI measurements, factors such as age-related differences in SNR become inherent to the error estimates that are incorporated into reliability calculations. While substantial progress has been made in understanding motion-related confounds in cross-sectional and between-subject fMRI analyses, additional research is needed to determine how to best address the effects of within-subject changes in motion and their subsequent effects on fMRI test-retest reliability estimates.

4.4. Within- and between-subject registration of brain activation

Alignment of structural and functional images is imperative to ensuring meaningful within- and between-subject comparisons of brain activity. Specifically, poor alignment of brain regions will certainly lead to poor test-retest reliability, especially when examining ICC values at the voxelwise level or in smaller regions of interest. Importantly, common registration techniques for fMRI data are largely based on registration of structural brain MRI scans. During childhood and adolescence, brain structure, including whole-brain volume, cortical thickness, and sulcal topology undergo significant maturation (Vandekar et al., 2015; Mills et al., 2016; Tamnes et al., 2017). Similarly, standard-preprocessing steps for functional data, such as the Gaussian smoothing kernel, can influence ICC values. For example, by increasing a smoothing kernel, both between-subject variance and error variance may decrease, which in turn may increase ICC estimates (Caceres et al., 2009). Optimizing subject registration of structural and functional data across timepoints is therefore vital to estimating accurate test-retest reliability metrics in developing samples.

4.5. Statistical approaches to disentangle true developmental change and reliability error

Given the number of confounding factors mentioned above that may directly influence the fMRI BOLD signal over time, researchers may consider adopting a more complex study design and incorporating additional statistical approaches to 1) establish test-retest reliability estimates for a given fMRI task paradigm in children and adolescents and 2) to also correct for reliability error during statistical modeling of true developmental change. Specifically, it would be useful for future research to compare test-retest reliability of a task within a subsample using a short time interval (days or weeks) that is unlikely to be confounded by large developmental changes between measurements. After establishing reliability with a short time interval, these estimates can be

used to correct error estimations of the fMRI BOLD signal in typical developmental studies with longer intervals (e.g. 1 or 2 years). In other words, for a given fMRI task paradigm, one may establish the test-retest reliability of the fMRI BOLD signal in a replication sub-sample of children or adolescents from the larger study sample. This fMRI data can then be used to derive the test-retest reliability of each task condition and for various brain ROIs. Using these newly established ICC estimates, a bias-correction formula can be generated to calibrate the longitudinal estimates of BOLD signal changes across development for a given fMRI task condition or brain ROI. This type of measurement error bias-correction approach has been widely used in psychological research to determine significant changes while controlling for the contribution of reliability of the metric. Similarly, by examining the ICC values of the BOLD signal for a given fMRI task condition, brain region (voxel or ROI), and population (age of interest), the random measurement error due to test-retest reliability can be estimated from the replication sub-sample and accounted for in the longitudinal analysis. For example, for a longitudinal fMRI study in children and adolescents that performs only a single baseline and one follow-up timepoint, a linear regression using the change model can be used to assess the relationship between baseline and change between the two timepoints (Blmoqvist, 1977). A bias-correction formula including the reliability ratio can then be applied to improve the estimates of the BOLD signal to better estimate the relationship between BOLD signal at baseline and changes in BOLD signal over time. For multiple measurement occasions, including polynomial functions, measurement error in the repeated-measures can be estimated by the residual error of mixed effects models; as such, bias-correction formulas have also been developed for analyzing the associations between multiple factors and the longitudinal trajectory (Byth and Cox, 2005; Harrison et al., 2009; also see (Chen, 2013) for more detail).

5. Conclusions and future directions

When designing a longitudinal study, the most obvious recommendation is to choose fMRI tasks with high reliability based on your sample and question of interest. Moving forward, all task-based fMRI longitudinal studies should report ICC values, including the type of ICC and the confidence intervals for each ICC value, as part of their results. As the field progresses, it will become increasingly important to build a reference library of comparative reliabilities for different fMRI tasks across age groups. Although not all previously published longitudinal task-based fMRI studies in children and adolescents have reported ICC values, it would still be extremely useful for previous developmental task-based fMRI studies to provide ICC values for whole-brain or multiple ROIs to the scientific community and/or openly share their data through common data sharing platforms (i.e. https://openfmri.org (Poldrack et al., 2013)). By building a reference library of test-retest reliability reports, future studies will be better equipped to consider the reliability of the spatial location and BOLD signal for *a priori* ROIs during the study design phase. Of course, as newly designed fMRI task paradigms are being implemented in specific populations, then pilot studies will be necessary to estimate the ICC values of ROIs for the given sample (children, adolescents, with and without a clinical disorder) prior to study initiation. Ongoing and future longitudinal fMRI studies should also consider conducting reliability validation using sub-samples to correct for reliability error bias for the fMRI BOLD signal for a given task contrast and brain region. Overall, this is an exciting time for developmental neuroimaging as prospective longitudinal designs are becoming more widespread and will allow for greater insight into within- and between-subject differences in child and adolescent neurodevelopment. By considering the issues discussed in this review and taking steps to reduce test-retest reliability error estimates of BOLD signal, we will be able to more accurately elucidate true developmental change in brain activity using task-based fMRI.

Funding

This work was supported by the National Institute of Health [K01 MH108761 (Herting)] and by the German Ministry of Education and Research (BMBF grants # 01 EV 0711, 01 EE 1406B), the Deutsche Forschungsgemeinschaft (SFB 940/1, VE 892/2-1), and the MedDrive Start Grant of the Medical Faculty of the Technische Universität Dresden.

Conflict of Interest

None.

References

ABCD, 2015. Adolescent Brain Cognitive Development Study. (Retrieved January 15, 2016, from http://addictionresearch.nih.gov/adolescent-brain-cognitive-development-study).
Aron, A.R., Gluck, M.A., Poldrack, R.A., 2006. Long-term test-retest reliability of functional MRI in a classification learning task. Neuroimage 29 (3), 1000–1006.
Bartlett, J.W., Frost, C., 2008. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound Obstet. Gynecol. 31 (4), 466–475.
Beglinger, L.J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D.A., Crawford, J., Fastenau, P.S., Siemers, E.R., 2005. Practice effects and the use of alternate forms in serial neuropsychological testing. Arch. Clin. Neuropsychol. 20 (4), 517–529.
Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci. 1191, 133–155.
Bennett, C.M., Miller, M.B., 2013. fMRI reliability: influences of task and experimental design. Cogn. Affect Behav. Neurosci. 13 (4), 690–702.
Blakemore, S.J., 2012. Imaging brain development: the adolescent brain. Neuroimage 61 (2), 397–406.
Blumenthal, J.D., Zijdenbos, A., Molloy, E., Giedd, J.N., 2002. Motion artifact in magnetic resonance imaging: implications for automated analysis. Neuroimage 16 (1), 89–92.
Braams, B.R., Crone, E.A., 2016. Longitudinal changes in social brain development: processing outcomes for friend and self. Child Dev. http://dx.doi.org/10.1111/cdev.1266.
Braams, B.R., van Duijvenvoorde, A.C., Peper, J.S., Crone, E.A., 2015. Longitudinal changes in adolescent risk-taking: a comprehensive study of neural responses to rewards, pubertal development, and risk-taking behavior. J. Neurosci. 35 (18), 7226–7238.
Brandt, D.J., Sommer, J., Krach, S., Bedenbender, J., Kircher, T., Paulus, F.M., Jansen, A., 2013. Test-retest reliability of fMRI brain activity during memory encoding. Front Psychiatry 4, 163.
Britton, J.C., Bar-Haim, Y., Clementi, M.A., Sankin, L.S., Chen, G., Shechner, T., Norcross, M.A., Spiro, C.N., Lindstrom, K.M., Pine, D.S., 2013. Training-associated changes and stability of attention bias in youth: implications for Attention Bias Modification Treatment for pediatric anxiety. Dev. Cogn. Neurosci. 4, 52–64.
Byth, K., Cox, D.R., 2005. On the relation between initial value and slope. Biostatistics 6 (3), 395–403.
Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage 45 (3), 758–768.
Casey, B.J., Getz, S., Galvan, A., 2008. The adolescent brain. Dev. Rev. 28 (1), 62–77.
Chen, Z., 2013. Evaluating the Associations Between the Baseline and Other Exposure Variables with the Longitudinal Trajectory when Responses Are Measured with Error. PhD Dissertation. University of Southern California.
Church, J.A., Petersen, S.E., Schlaggar, B.L., 2010. The Task B problem and other considerations in developmental functional neuroimaging. Hum. Brain Mapp. 31 (6), 852–862.
Cicchetti, D.V., Sparrow, S.A., 1981. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am. J. Ment. Defic. 86 (2), 127–137.
Cicchetti, D.V., 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 (5), 695–700.
Crone, E.A., Dahl, R.E., 2012. Understanding adolescence as a period of social-affective engagement and goal flexibility. Nat. Rev. Neurosci. 13 (9), 636–650.
Crone, E.A., Elzinga, B.M., 2015. 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 (1), 53–63.
Dubois, J., Adolphs, R., 2016. Building a science of individual differences from fMRI. Trends Cogn. Sci. 20 (6), 425–443.
Eaton, K.P., Szaflarski, J.P., Altaye, M., Ball, A.L., Kissela, B.M., Banks, C., Holland, S.K., 2008. Reliability of fMRI for studies of language in post-stroke aphasia subjects. Neuroimage 41 (2), 311–322.
Forsyth, J.K., McEwen, S.C., Gee, D.G., Bearden, C.E., Addington, J., Goodyear, B., Cadenhead, K.S., Mirzakhanian, H., Cornblatt, B.A., Olvet, D.M., Mathalon, D.H., McGlashan, T.H., Perkins, D.O., Belger, A., Seidman, L.J., Thermenos, H.W., Tsuang, M.T., van Erp, T.G., Walker, E.F., Hamann, S., Woods, S.W., Qiu, M., Cannon, T.D., 2014. Reliability of functional magnetic resonance imaging activation during working memory in a multi-site study: analysis from the North American Prodrome

Longitudinal Study. Neuroimage 97, 41–52.
Genovese, C.R., Noll, D.C., Eddy, W.F., 1997. Estimating test-retest reliability in functional MR imaging. I: Statistical methodology. Magn. Reson. Med. 38 (3), 497–507.
Gorgolewski, K.J., Storkey, A., Bastin, M.E., Whittle, I.R., Wardlaw, J.M., Pernet, C.R., 2013a. A test-retest fMRI dataset for motor, language and spatial attention functions. Gigascience 2 (1), 6.
Gorgolewski, K.J., Storkey, A.J., Bastin, M.E., Whittle, I., Pernet, C., 2013b. Single subject fMRI test-retest reliability metrics and confounding factors. Neuroimage 69, 231–243.
Harrison, L., Dunn, D.T., Green, H., Copas, A.J., 2009. Modelling the association between patient characteristics and the change over time in a disease measure using observational cohort data. Stat. Med. 28 (26), 3260–3275.
Huettel, S.A., Song, A.W., McCarthy, G., 2004. Functional Magnetic Resonance Imaging. Sinauer Associates.
Hunt, R.J., 1986. Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability. J. Dent. Res. 65 (2), 128–130.
IMAGEN, 2007. IMAGEN Study. (Retrieved 15 January 2016, from http://www.imagen-europe.com/).
Khoo, S.T., West, S.G., Wu, W., Kwok, O.M., 2006. Longitudinal Methods. Handbook of Multimethod Measurement in Psychology. American Psychology Association.
Koolschijn, P.C., Schel, M.A., de Rooij, M., Rombouts, S.A., Crone, E.A., 2011. A three-year longitudinal functional magnetic resonance imaging study of performance monitoring and test-retest reliability from childhood to early adulthood. J. Neurosci. 31 (11), 4204–4212.
Lipp, I., Murphy, K., Wise, R.G., Caseras, X., 2014. Understanding the contribution of neural and physiological signal variation to the low repeatability of emotion-induced BOLD responses. Neuroimage 86, 335–342.
McArdle, J.J., Woodcock, R.W., 1997. Expanding test-retest designs to include developmental time-lag components. Psychol. Methods 2 (4), 403–435.
McCormick, E.M., Qu, Y., Telzer, E.H., 2016. Adolescent neurodevelopment of cognitive control and risk-taking in negative family contexts. Neuroimage 124 (Pt A), 989–996.
Mills, K.L., Goddings, A.L., Herting, M.M., Meuwese, R., Blakemore, S.J., Crone, E.A., Dahl, R.E., Guroglu, B., Raznahan, A., Sowell, E.R., Tamnes, C.K., 2016. Structural brain development between childhood and adulthood: convergence across four longitudinal samples. Neuroimage 141, 273–281.
NCANDA, 2014. National Consortium on Alcohol & Neurodevelopment in Adolescence. (Retrieved 15 January 2016, from http://www.ncanda.org/).
Ordaz, S.J., Foran, W., Velanova, K., Luna, B., 2013. Longitudinal growth curves of brain function underlying inhibitory control through adolescence. J. Neurosci. 33 (46), 18109–18124.
Paulsen, D.J., Hallquist, M.N., Geier, C.F., Luna, B., 2015. Effects of incentives, age, and behavior on brain activation during inhibitory control: a longitudinal fMRI study. Dev. Cogn. Neurosci. 11, 105–115.
Peters, S., Van Duijvenvoorde, A.C., Koolschijn, P.C., Crone, E.A., 2016. Longitudinal development of frontoparietal activity during feedback learning: contributions of age, performance, working memory and cortical thickness. Dev. Cogn. Neurosci. 19, 211–222.
Plichta, M.M., Grimm, O., Morgen, K., Mier, D., Sauer, C., Haddad, L., Tost, H., Esslinger, C., Kirsch, P., Schwarz, A.J., Meyer-Lindenberg, A., 2014. Amygdala habituation: a reliable fMRI phenotype. Neuroimage 103, 383–390.
Poldrack, R.A., Barch, D.M., Mitchell, J.P., Wager, T.D., Wagner, A.D., Devlin, J.T., Cumba, C., Koyejo, O., Milham, M.P., 2013. Toward open sharing of task-based fMRI data: the OpenfMRI project. Front Neuroinform 7, 12.
Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2012. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 59 (3), 2142–2154.
Qu, Y., Fuligni, A.J., Galvan, A., Telzer, E.H., 2015. Buffering effect of positive parent-child relationships on adolescent risk taking: a longitudinal neuroimaging investigation. Dev. Cogn. Neurosci. 15, 26–34.
Schlaggar, B.L., Brown, T.T., Lugar, H.M., Visscher, K.M., Miezin, F.M., Petersen, S.E., 2002. Functional neuroanatomical differences between adults and school-age children in the processing of single words. Science 296 (5572), 1476–1479.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 (2), 420–428.
Singer, J.D., Willett, J.B., 2003. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. Oxford University Press, Inc, New York.
Tamnes, C.K., Herting, M.M., Goddings, A.L., Meuwese, R., Blakemore, S.J., Dahl, R.E., Guroglu, B., Raznahan, A., Sowell, E.R., Crone, E.A., Mills, K.L., 2017. Development of the cerebral cortex across adolescence: a multisample study of inter-related longitudinal changes in cortical volume, surface area, and thickness. J. Neurosci. 37 (12), 3402–3412.
Vandekar, S.N., Shinohara, R.T., Raznahan, A., Roalf, D.R., Ross, M., DeLeo, N., Ruparel, K., Verma, R., Wolf, D.H., Gur, R.C., Gur, R.E., Satterthwaite, T.D., 2015. Topologically dissociable patterns of development of the human cerebral cortex. J. Neurosci. 35 (2), 599–609.
van den Bulk, B.G., Koolschijn, P.C., Meens, P.H., van Lang, N.D., van der Wee, N.J., Rombouts, S.A., Vermeiren, R.R., Crone, E.A., 2013. 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.
Vetter, N.C., Pilhatsch, M., Weigelt, S., Ripke, S., Smolka, M.N., 2015. Mid-adolescent neurocognitive development of ignoring and attending emotional stimuli. Dev. Cogn. Neurosci. 14, 23–31.
Vetter, N.C., Steding, J., Jurk, S., Ripke, S., Mennigen, E., Smolka, M.N., 2017. Reliability in adolescent fMRI within two years – a comparison of three tasks. Sci. Rep. 7 (1), 2287.
White, L.K., Britton, J.C., Sequeira, S., Ronkin, E.G., Chen, G., Bar-Haim, Y., Shechner, T., Ernst, M., Fox, N.A., Leibenluft, E., Pine, D.S., 2016. Behavioral and neural stability of attention bias to threat in healthy adolescents. Neuroimage 136, 84–93.
Willett, J.B., 1989. Some results on reliability for the longitudinal measurement of change: implications for the design of studies of individual growth. Educ. Psychol. Meas. 49, 587–602.