Developmental Cognitive Neuroscience 31 (2018) 67–73

Contents lists available at ScienceDirect

Developmental Cognitive Neuroscience

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

Reliability of neural activation and connectivity during implicit face emotion processing in youth

Simone P. Haller[a,*,1], Katharina Kircanski[a,1], Joel Stoddard[b], Lauren K. White[c], Gang Chen[d], Banafsheh Sharif-Askary[e], Susan Zhang[a], Kenneth E. Towbin[a], Daniel S. Pine[a], Ellen Leibenluft[a], Melissa A. Brotman[a]

[a] Emotion and Development Branch, National Institute of Mental Health, USA
[b] Department of Psychiatry, University of Colorado School of Medicine, USA
[c] Children's Hospital of Philadelphia, Lifespan Brain Institute, USA
[d] Scientific and Statistical Computing Core, National Institute of Mental Health, National Institutes of Health, USA
[e] Duke University, School of Medicine, USA

ARTICLE INFO

Keywords:
Reliability
fMRI
Emotion processing
Children and adolescents

ABSTRACT

Face emotion imaging paradigms are widely used in both healthy and psychiatric populations. Here, in children and adolescents, we evaluate the test-retest reliability of blood oxygenation-level dependent (BOLD) activation and task-based functional connectivity on a widely used implicit face emotion processing task (i.e., gender labeling). Twenty-five healthy youth (*M* age = 13.97 years; 60% female) completed two functional magnetic resonance imaging (fMRI) scan sessions approximately two months apart. Participants identified the gender of faces displaying angry, fearful, happy, and neutral emotions. A Bayesian adaptation of the intraclass correlation (ICC) assessed reliability of evoked BOLD activation and amygdala seed-based functional connectivity on task events vs. baseline as well as contrasts between face emotions. For each face emotion vs. baseline, good reliability of activation was demonstrated across key emotion processing regions including middle, medial, and inferior frontal gyri. However, contrasts between face emotions yielded variable results. Contrasts of angry to neutral or happy faces exhibited good reliability of amygdala connectivity to prefrontal regions. Contrasts of fearful to happy faces exhibited good reliability of activation in the anterior cingulate. Findings inform the reproducibility literature and emphasize the need for continued evaluation of task reliability.

1. Introduction

A growing body of work examines the test-retest reliability of functional magnetic resonance imaging (fMRI) measures (reviewed in Bennett and Miller, 2010). This critical research area addresses efforts to improve the reproducibility of findings in biomedical research generally (e.g., Collins and Tabak, 2014). The current study assesses the reliability of activity and amygdala task-based functional connectivity in healthy youth for a commonly-used implicit face emotion processing paradigm.

Facial displays of emotion are among the most common stimuli used to study emotion processing across healthy and psychiatric populations. One of the most frequently employed face emotion paradigms probes implicit processing by focusing participants' attention on a non-emotional feature of the face, such as gender. This task robustly activates the amygdala and various regions in the prefrontal cortex (PFC; e.g., Nord et al., 2017; Stoddard et al., 2017) central to emotion processing (e.g., Etkin et al., 2011; Kober et al., 2008). Findings broadly inform understanding of various psychiatric conditions in youth and adults (Fonville et al., 2014; Hassel et al., 2009; Kalmar et al., 2009; Kim et al., 2012; Lawrence et al., 2004; Rosenfeld et al., 2014; Shah et al., 2008; Surguladze et al., 2005, 2010; Thomas et al., 2013). Most recently, work employing this paradigm has shown that youth with clinical levels of chronic irritability and anxiety exhibit perturbations in fronto-amygdala connectivity to intensely angry faces (Stoddard et al., 2017). These and related findings from such implicit face viewing tasks vitally inform neuroscience research on multiple pediatric psychiatric phenotypes.

However, robust activation in task contrasts when averaging across individuals is distinct from *reliably* evoked BOLD activation *at the level of the individual* (Nord et al., 2017). Hence, the reliability of one individual's level of activation or connectivity cannot be inferred from

* Corresponding author at: Emotion and Development Branch, National Institute of Mental Health, 9000 Rockville Pike, Building 15K, MSC-2670, Bethesda, MD 20892-2670, USA.
E-mail address: simone.haller@nih.gov (S.P. Haller).
[1] These authors contributed equally.

https://doi.org/10.1016/j.dcn.2018.03.010
Received 22 September 2017; Received in revised form 28 February 2018; Accepted 24 March 2018
Available online 27 March 2018

group level activation at a single point in time. Compared to adults, very little is known about reliability of task-based fMRI measures in youth. Investigating reliability in youth is particularly important given the reliance of large-scale longitudinal developmental work on task-based fMRI (e.g., ABCD). In addition, there are specific concerns about in-scanner motion in children; reliability estimates derived from adult work may not hold for developmental populations.

Recent imaging studies quantifying the consistency of individual rankings over time rely on the intraclass correlation coefficient (ICC) (Caceres et al., 2009; Shrout and Fleiss, 1979). Several studies use this measure to evaluate task-evoked measures of the blood oxygenation-level dependent (BOLD) signal (e.g., Bennett and Miller, 2010), and one study reports stability of functional connectivity (White et al. 2016). A few such investigations rely on face processing paradigms, including tasks requiring face matching (e.g., Nord et al., 2017; Sauder et al., 2013; Plichta et al., 2012), labeling (Nord et al., 2017), attention orienting (i.e., dot-probe; Britton et al., 2013; White et al., 2016), and passive viewing (e.g., Johnstone et al., 2005). Only three studies have examined reliability in pediatric samples. These studies reported divergent estimates for both frontal and amygdala activation, ranging from "fair" to "good" in frontal activation and "poor" to "good" in amygdala activation across studies (Britton et al., 2013; van den Bulk et al., 2013; White et al., 2016). Adult studies largely reported "moderate" to "good" estimates for activation in PFC regions when face emotions were compared to an implicit baseline. However, similar to the pediatric work, reliability estimates were highly variable (Bunford et al., 2017; Johnstone et al., 2005; Manuck et al., 2007; Plichta et al., 2012; Sauder et al., 2013).

Most previous work has only examined reliability of regional BOLD activation, yet task-based functional connectivity is an increasing focus of research on psychopathology and development. Psychophysiological interaction (PPI) analyses is a method for the investigation of task-related changes in the relationship between activation in a region of interest and other brain areas. Only one study to date has investigated both neural activation and task-based functional connectivity on a fMRI paradigm relevant to psychopathology research. The authors reported consistently higher reliability estimates for fronto-amygdala connectivity than for regional PFC activation in a pediatric sample in an attention orienting task with face emotions (White et al., 2016). Fronto-amygdala circuitry is pertinent to both normative and aberrant emotion processing and regulation during the developmental period of late childhood and adolescence (e.g., Gee et al., 2013; Stoddard et al., 2017). Thus, these first data on reliability of connectivity in this circuit are encouraging, but require extension to other widely used face emotion paradigms.

1.1. The current study

We evaluated the test-retest reliability of neural activation and functional connectivity on an implicit face emotion processing paradigm in children and adolescents. Participants identified the gender of faces displaying three emotions (angry, fearful and happy) at three intensities of expression (50%, 100% and 150%), interspersed with affectively neutral faces. We derived reliability estimates across two MRI sessions using a Bayesian method of the ICC. Reliability was assessed for an *a priori* series of eight analytic contrasts, based on their prevalence in the literature. Specifically, we focused on the more widely applicable aspect of the paradigm, the emotion contrasts. First, we examined the reliability of activation for each emotion (i.e., happy, angry, fearful) as compared to implicit baseline (i.e., fixation). Second, we examined the reliability of specific contrasts between emotions for both activation and functional connectivity: each emotion (i.e., happy, angry, fearful) relative to neutral, and each negatively valenced emotion (i.e., angry, fearful) relative to the one positive emotion (i.e., happy). Based on prior studies, we expected contrasts of each emotion to baseline to show higher reliability than contrasts between emotions. Given previous work by White et al. (2016), we expected higher reliability estimates for fronto-amygdala connectivity than activation.

2. Methods

2.1. Participants

Participants were 25 healthy children and adolescents (*M* age = 13.97 years, *SD* = 2.22, range = 10.04–17.51; 60% female; *M* Tanner Stage = 3.9, SD = 0.93 [n = 18]). Five additional participants were enrolled in the study but were excluded due to poor behavioral performance ($n = 1$), technical difficulties during scanning ($n = 1$), or excess motion ($n = 3$). All participants completed a structured clinical interview (K-SADS-PL; Kaufman et al., 1997) confirming that they had no current or lifetime DSM-IV mental disorders. All participants had an IQ > 70 as assessed using the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999). Signed parental consent and youth assent were obtained prior to participation. The study protocol was approved by the National Institute of Mental Health Institutional Review Board.

2.2. Task

Participants completed two MRI scanning sessions approximately two-and-a-half months apart (*M* = 75.12 days, *SD* = 15.12, range: 47-109). During each functional MRI scan, participants labeled the gender (male, female) of each of 10 actors' faces portraying angry, fearful, happy and neutral expressions (Ekman and Friesen, 1976). Each face emotion (happy, angry, fearful) was presented at three intensities of expression (50%, 100%, 150%) across trials. Face emotion stimuli were presented in randomized order for 2000 ms (ms), followed by a jittered fixation (*M* = 1400 ms; range = 500–6000, see Fig. 1). Trials were presented in three blocks, including a total of 30 trials of each face emotion at each intensity and 90 neutral trials. The task was administered in E-Prime (Psychology Software Tools, USA). The task performance criterion was > 65% gender identification across all face stimuli (Stoddard et al., 2017).

2.3. fMRI data acquisition

Magnetic resonance images were acquired on a General Electric 3 T

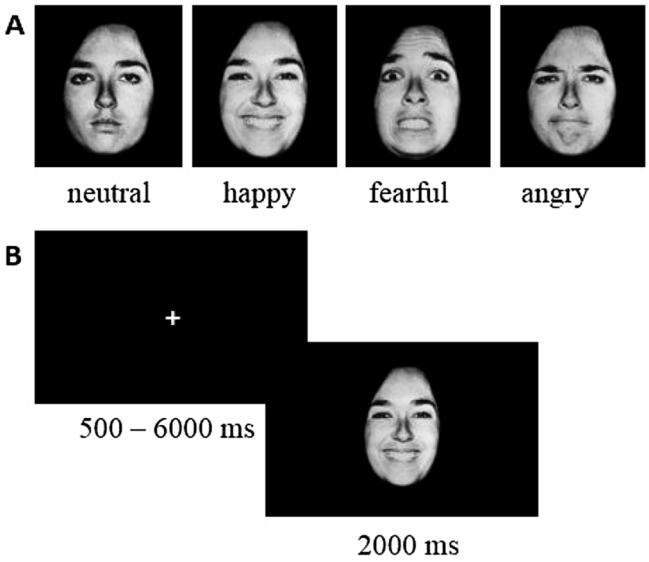

Fig. 1. (A) Example stimuli. Faces were drawn from the Ekman and Friesen (1976) standardized face set. (B) Example trial sequence. Stimuli were presented for 2000 ms, followed by a jittered fixation.

scanner (Waukesha, WI, USA) with a 32-channel head coil. BOLD signal was measured by echoplanar imaging at a voxel resolution of 2.5 × 2.5 × 3.0 mm (flip angle = 50°, repetition time = 2300 ms, echo time = 25 ms, field of view = 240 mm). Total acquisition time was 21 minute s; trials were split across three runs with 182 volumes per run.

For the purpose of co-registration and normalization, T1-weighted magnetization-prepared 180° radio-frequency pulses and rapid gradient-echo (MPRAGE) images were collected (flip angle = 7°, minimum full echo time, inversion time = 425 ms, acquisition voxel size = 1 mm isotropic).

2.4. fMRI data processing

fMRI data were pre-processed using Freesurfer (Ségonne et al., 2004) and Analysis of Functional Neuroimages (AFNI; Cox, 1996) standard software. MPRAGE images were skull stripped with Freesurfer. Pre-processing included slice-timing correction, alignment and non-linear registration to a Talaraich template, spatial smoothing (5-mm full-width half-maximum kernel) and scaling each voxel to a run mean of 100. Individual TRs and the immediately preceding volume were censored if: (1) the motion shift, defined as Euclidean norm of the derivative of the translation and rotation parameters, exceeded 1 mm between TRs; and (2) more than 10% of voxels were outliers. Participants were excluded if the average motion per TR after censoring was > 0.25 mm or if more than 15% of TRs were censored for motion/outliers. An average of 3.3% (*SD* = 3.2%) of TRs were censored at T1, and 2.9% (*SD* = 3.2%) at T2.

2.5. Statistical analyses

Analyses used AFNI version 16.2.16 and RStudio version 3.4.

2.5.1. BOLD activation

A general linear model (GLM) was created for each participant at each time point. Only trials with accurate gender identification were included for effects of interest, but incorrect trials were also modelled. Separate regressors were created for each of 13 event types (i.e., each face emotion at each intensity, neutral trials represented by three regressors of 30 trials each, and an additional regressor coding incorrect trials). Additional regressors included six head motion parameters and baseline drift using third order Legendre polynomials. Individual level contrasts were then submitted to the intraclass correlation (ICC) analyses.

2.5.2. Functional connectivity

Functional connectivity was analyzed using generalized psychophysiological interaction (gPPI) methods (McLaren et al., 2012). The right and left amygdala were used as seed regions, defined anatomically using the Talaraich atlas and using only voxels within which 90% or more of all participants had data for both sessions. Separate GLMs were created for each seed region.

PPI regressors were created as the product of the detrended and demeaned seed regressor and the psychophysiological event. The additional 13 PPI regressors and the seed's mean time series were then added to the individual level GLMs, identical to the GLMs used to estimate neural activation with motion and drift parameters, and including identical general linear tests for ICC contrasts of interest.

2.5.3. Test-retest reliability analyses

A Bayesian intraclass correlation (ICC) approach (Chen et al., 2018) was used to compute voxel-wise ICCs of BOLD activation and amygdala seed-based functional connectivity across the two sessions. We used the Bayesian ICC approach as it has been demonstrated to address potential issues in traditional ICC estimates (e.g., negative ICC values, missing data, confounding effects). ICCs with absolute agreement (ICC (2,1); Shrout and Fleiss, 1979) were modeled with Gamma priors for the random variables of 'subject' and 'visit' in a Linear Mixed-Effects model (3dLME; Chen et al., 2013). We used the absolute ICC (2,1) model as it reflects reliability based on absolute agreement, which applies a more stringent criterion than does a consistency model. Further, the only other paper to examine reliability of neural activation and functional connectivity in a pediatric sample (White et al., 2016) utilized the absolute ICC, which makes our findings more directly comparable to that previous study. We submitted the individual level contrasts from the BOLD and PPI GLMs to the ICC analyses.

We investigated three sets of contrasts. First, we examined the reliability evoked for each emotion, averaged across intensity, compared to baseline fixation. Next, we examined the reliability of contrasts between each emotion, averaged across intensity, and neutral faces. Last, we assessed the reliability of contrasts between positive emotion and each negative emotion (i.e., happy vs. angry, happy vs. fearful), averaged across intensity.

The ICC threshold was set to 0.50, corresponding to an uncorrected $p < .005$ threshold with 24 degrees of freedom (Bartko, 1966). As most previous research on face emotion processing focuses on frontal regions and fronto-limbic connectivity, ICC analyses were performed across a prefrontal cortex (PFC) gray matter mask i.e., 18,689 voxels, including only voxels for which 90% or more of participants had data. Masked results were thresholded at voxel-wise $p < .005$ with an overall family-wise error rate of $\alpha < .05$, yielding k = 42 contiguous voxels (calculated via AFNI's 3dClustSim; Monte-Carlo cluster size simulation with a Gaussian plus exponential spatial autocorrelation function to estimate smoothness; a = 0.54, b = 3.99, c = 10.86 mm FWHM derived from the data collected in this study) (see Cox et al., 2017).

For statistically significant clusters, AFNI's 3dROIstat was used to extract mean activity and connectivity during Session 1 and Session 2. Individual participant mean beta or PPI values considered outliers [i.e., greater than 3.29 *SD*s (99% CI)] were identified in each extracted cluster and excluded. Only those clusters that remained significant after the removal of outliers are presented. Clusters are reported with size (k), peak ICC value, and coordinates in Talaraich space.

We additionally conducted post-hoc region-of-interest (ROI) reliability analyses on amygdala activation across the same eight contrasts. We extracted activation for each session from the right and left amygdala based on the anatomically defined amygdala mask used as a seed in the PPI analyses. ICC values across ROIs for each contrast were calculated using R and the package *blme* (Chung et al., 2013). Additional analyses controlling for age and interval between sessions are in the Supplementary Material.

3. Results

3.1. Reliability of BOLD activation

Full results for areas with reliable neural activation are presented in Table 1. Our first analysis examined each face emotion (happy, angry, fearful) versus baseline. This revealed good reliability estimates for activation across many regions of the PFC, including the middle, inferior and medial frontal gyri (see Fig. 2 for example clusters across baseline contrasts). However, our second analysis examined contrasts of each face emotion versus neutral faces, and did not detect areas of stable neural activation. Our third set of analyses examined negative versus positive emotions, and this analysis found one contrast with one cluster passing our threshold for stable activation. Specifically, the contrast of happy versus fearful faces yielded a stable cluster in the right anterior cingulate cortex (ACC; see Fig. 3).

Within an amygdala ROI, we quantified reliability of activation across these same task contrasts. Overall, reliability was poor, ranging from 0.10 (angry vs. neutral) to 0.36 (fearful vs. happy) in the right amygdala, and 0.15 (angry vs. neutral) to 0.38 (fearful vs. happy) in the left amygdala. Full results for all eight contrasts for each ROI are in the

Table 1
Summary of reliability estimates for BOLD activation.

Task condition	Peak Talaraich Coordinates			Cluster size k	ICC peak value	Peak TLRC Location
	x	y	z			
Angry vs. BL	31	11	26	126	0.78	R Middle Frontal Gyrus
	54	26	21	50	0.64	R Inferior Frontal Gyrus
Fearful vs. BL	−51	26	24	302	0.77	L Inferior Frontal Gyrus
	34	11	26	248	0.87	R Middle Frontal Gyrus
	−29	11	56	66	0.73	L Middle Frontal Gyrus
Happy vs. BL	36	11	26	100	0.70	R Middle Frontal Gyrus
	1	49	1	58	0.66	R Medial Frontal Gyrus
	29	44	36	44	0.63	R Superior Frontal Gyrus
	−59	6	29	43	0.69	L Inferior Frontal Gyrus
Angry vs. Neutral	–	–	–	–	–	–
Fearful vs. Neutral	–	–	–	–	–	–
Happy vs. Neutral	–	–	–	–	–	–
Angry vs. Happy	–	–	–	–	–	–
Fearful vs. Happy	4	39	6	139	0.74	R Anterior Cingulate

Note: BL, Baseline; R, right; L, left.

Supplementary material.

3.2. Reliability of functional connectivity

Full results for the reliability of fronto-amygdala connectivity are presented in Table 2. Our first analysis examined differences in amygdala connectivity evoked by individual face emotions versus neutral faces. This revealed two significant clusters. Specifically, the contrast of angry versus neutral faces yielded stable changes in connectivity between the left amygdala and left middle frontal gyrus. The comparisons of negative versus positive emotions also revealed several significant clusters. Specifically, the contrast of happy versus angry faces yielded clusters reflecting stable changes in connectivity between the left amygdala and multiple frontal regions (see Fig. 4). No clusters were significant for connectivity of the right amygdala to any PFC regions.

4. Discussion

The aim of this study was to investigate the reliability of task-based activation and connectivity on an implicit face emotion paradigm in children and adolescents. Recent adult work (Nord et al., 2017) has found reliability of task-evoked BOLD response on common face emotion processing paradigms to be surprisingly poor for areas considered candidate biomarkers in psychopathology research. The reliability of fMRI measures determines whether they can be helpful in identifying neural substrates of normative emotion processing and potential biomarkers of psychopathology. Reliability provides critical information for adequate power estimation needed to promote reproducibility. A limited number of prior studies have investigated reliability of both activation and connectivity during face emotion processing, and only one has done so in a pediatric sample (White et al., 2016).

As predicted, good reliability of activation emerged across many regions when contrasting each face emotion to baseline. These findings were largely consistent across angry, fearful, and happy stimuli, suggesting that neural activity during implicit face emotion processing is broadly reliable relative to baseline. However, contrasts between neutral faces and face emotions yielded generally poor reliability. In fact, on contrasts of specific emotions with neutral faces, only measures of connectivity generated any results passing our threshold for meaningful reliability.

Both angry and fearful face emotions have been used to study threat processing. Whereas angry faces represent a direct social threat, fearful

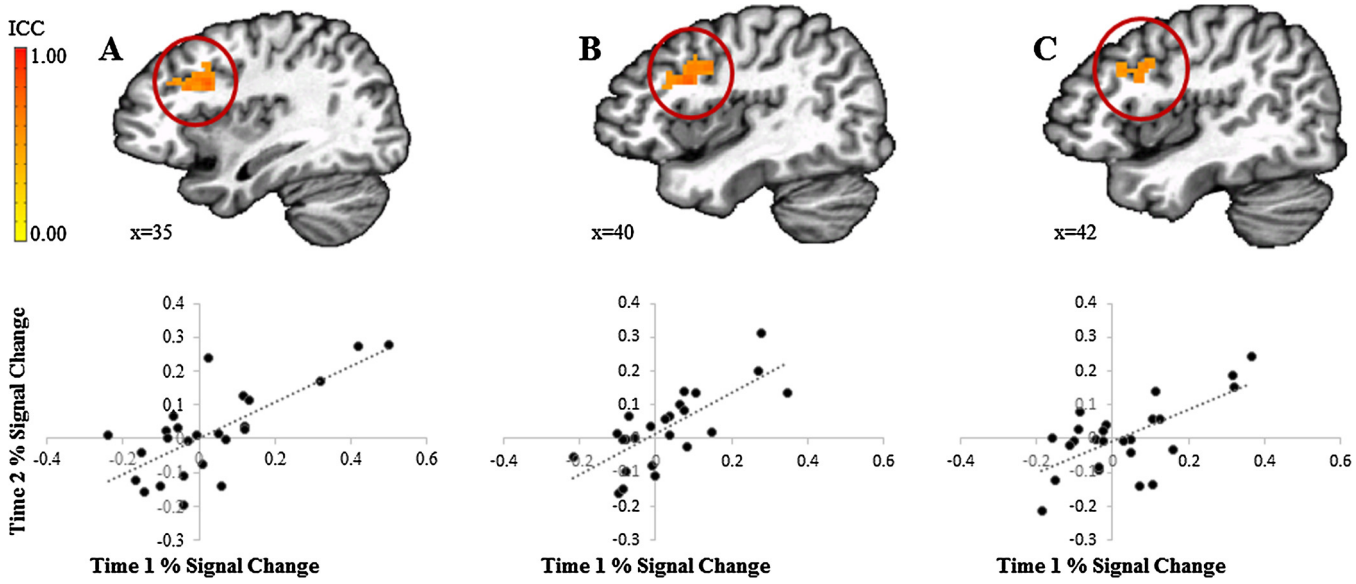

Fig. 2. Example clusters derived from comparing each emotion (angry, fearful, happy) to baseline. Right middle frontal gyrus exhibited reliable activation across all three contrasts. Graphs below each image plot BOLD (% signal change) values at each session from the associated activation cluster. Images show the following contrasts: (A) angry [k = 126 voxels], (B) fearful [k = 248 voxels], and (C) happy faces [k = 100 voxels] compared to baseline.

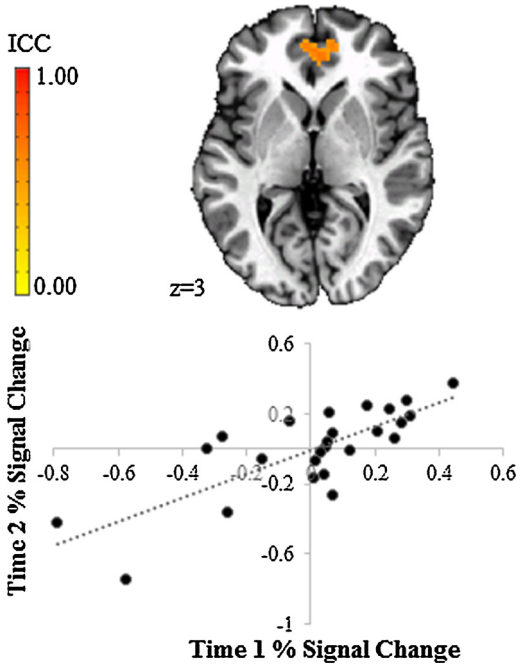

Fig. 3. Reliable activation in the ACC [k = 139] in the fearful versus happy contrast. Graph below the image plots BOLD (% signal change) values at each session from the associated activation cluster.

faces are thought to signal threat indirectly (i.e., presence of a threat in the environment) (Whalen et al., 2001). Neither stimulus produced significantly reliable activations relative to neutral face emotion, though other contrasts did reveal significant results. Specifically, activation in the ACC was reliable for the contrasts of fearful versus happy stimuli. Moreover, angry faces evoked reliable differences in connectivity. Specifically, differences in amygdala connectivity with various PFC regions were reliable for the contrast of angry versus happy or neutral stimuli. Such findings for connectivity estimates with angry faces generally replicate the pattern in a previous reliability study using a different task (White et al., 2016). This is a notable exception to the general pattern of poor reliability both in this study and other studies for emotion-specific contrasts.

These results highlight the importance of studying both activation and connectivity in relation to social threat processing. It is currently unclear why a larger number of reliable effects were found for connectivity than activation (see also White et al., 2016). Future research should continue to investigate reliability of both fMRI measures, which will demonstrate whether this pattern of findings for connectivity versus activation replicates. Following replication, examination of the reasons for this consistent pattern would be highly informative.

We did not find reliable activation in the contrasts of each emotion to canonically neutral faces. Past work demonstrates that such faces are not necessarily perceived as "neutral" (e.g., Brotman et al., 2009; Marusak et al., 2017), with factors such as state anxiety linking to variability in activation to neutral faces (Somerville et al., 2004). Previous work has discouraged the use of neutral faces as baseline conditions. Some suggest that neutral faces represent particularly poor baseline stimuli in youth (see Marusak et al., 2017; Thomas et al., 2007), the age targeted in our study. Variability and context dependency in perceiving the neutral face emotion stimuli may also contribute to poor reliability. Our data raise further questions about suitable baseline conditions for studies of face emotion processing.

Additional ICC analyses of the amygdala indicated low overall reliability of activation across task conditions. The highest ICC values, 0.36–0.38 bilaterally, were found in the contrast of fearful versus happy stimuli. Previous studies of the reliability of amygdala activation have yielded mixed results; one study reported good estimates (Britton et al., 2013), whereas two other studies reported poor temporal stability of amygdala activation (i.e., ICCs < .4, e.g., van den Bulk et al., 2013; White et al., 2016; Nord et al., 2017). Task design (e.g., number of repetitions, number of face emotions) and preprocessing strategies likely play a role in these variable estimates. The preprocessing used in the current paper is based on common standards such that the results may be relevant to a large portion of the field using this paradigm.

4.1. Strengths and limitations

This study has several strengths. It represents one of a few efforts to examine test-retest reliability in an fMRI paradigm on emotion processing, especially in youth. It is only the second study in the literature to examine reliability of both activation and task-based connectivity. Examining reliability of this widely used implicit emotion processing paradigm is critical, as the paradigm figures prominently in the field's ongoing search for biomarkers of psychopathology.

The results of this study also should be interpreted with caution in light of several limitations. First, although the sample size is comparable to that in many pediatric neuroimaging studies, it is not large. This increases sensitivity to individual outliers and generates imprecise estimates for reliabilities lying above chance but falling in the low-to-moderate range. Additionally, small sample sizes in combination with poor reliability limit the study's power to assess associations between brain function and inter-individual differences such as age. Additional analyses controlling for age (see Supplementary material) showed that the overall pattern of results held when controlling for age. However, given that this paradigm is often used in developmental research (e.g.,

Table 2
Summary of reliability estimates for reliability of fronto-amygdala connectivity.

Functional connectivity	Peak Talaraich Coordinates			Cluster size k	ICC peak value	Peak TLRC Location
	x	y	z			
Left Amygdala Seed						
Angry vs. Neutral	−39	11	41	99	.70	L Middle Frontal Gyrus
	−41	39	−6	64	.76	L Middle Frontal Gyrus
Fearful vs. Neutral	–	–	–	–	–	–
Happy vs. Neutral	–	–	–	–	–	–
Angry vs. Happy	−44	19	29	68	.69	L Middle Frontal Gyrus
	−46	31	−11	56	.70	L Inferior Frontal Gyrus
	−1	39	41	54	.67	R Medial Frontal Gyrus
	−6	24	49	50	.73	L Medial Frontal Gyrus
Fearful vs. Happy	–	–	–	–	–	–
Right Amygdala Seed						
	–	–	–	–	–	–

Note: BL, Baseline; R, right; L, left.

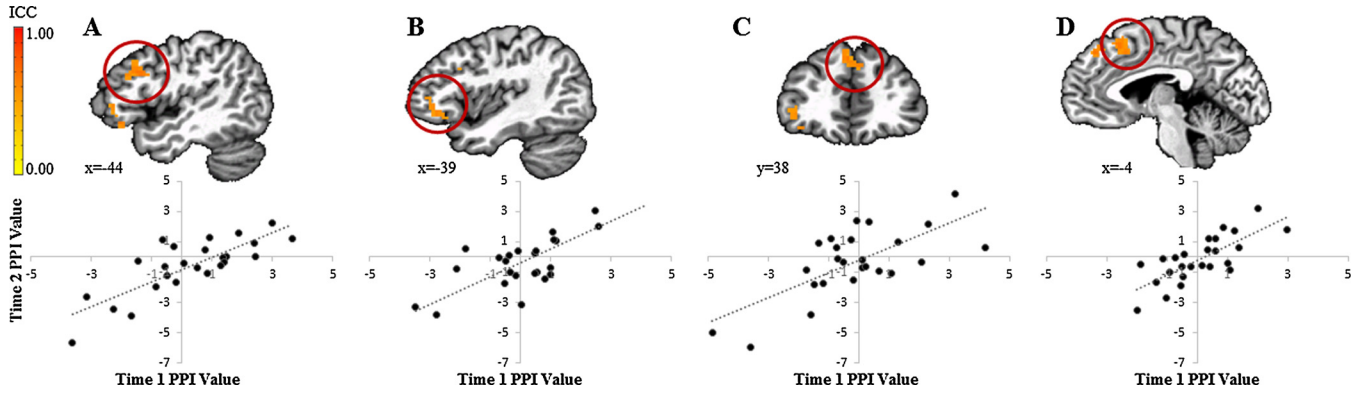

Fig. 4. Stability of fronto-amygdala connectivity for the angry versus happy contrast. Graphs below each image plot PPI values at each visit from the associated activation cluster. Results show stable connectivity between the left amygdala and several regions of the PFC, including: (A) left middle frontal [k = 68 voxels], (B) left inferior frontal [k = 56 voxels], (C) right medial frontal [k = 54 voxels], and (D) left medial frontal gyri [k = 50 voxels].

Garrett et al., 2015; Kalmar et al., 2009; Stoddard et al., 2017) and developmental effects could have affected ICC estimates, it will be important to assess this question more thoroughly. Second, this study investigated reliability in a healthy sample. While this helps establish a benchmark for reliability estimates, reliability may be different in clinical samples. As fMRI is used to help identify biomarkers of psychiatric illness, its reliability in clinical populations should be examined in future work.

4.2. Conclusion

We observed robust, reliable activation to facial stimuli relative to baseline. However, for contrasts utilizing neutral faces as a comparison condition, only measures of connectivity evoked by angry faces exhibited acceptable reliability. These results emphasize the need for continued evaluation of task reliability when trying to generate replicable results linking emotion-related brain function to various individual difference measures.

Conflict of interest

The authors declare no conflicts of interest.

Acknowledgements

This research was supported by the Intramural Research Program (IRP) of the National Institute of Mental Health, National Institutes of Health (NIMH/NIH), Grant Numbers ZIAMH002781 (Pine), ZIAMH002786 (Leibenluft) and ZIAMH002778 (Leibenluft).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.dcn.2018.03.010.

References

Bartko, J., 1966. The intraclass correlation coefficient as a measure of reliability. Psychol. Rep. 19, 3–11.
Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N.Y. Acad. Sci. 1191 (1), 133–155.
Britton, J.C., Bar-Haim, Y., Clementi, M.A., Sankin, L.S., Chen, G., Shechner, T., Pine, D.S., et al., 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. http://dx.doi.org/10.1016/j.dcn.2012.11.001.
Brotman, M.A., Rich, B.A., Guyer, A.E., Lunsford, J.R., Horsey, S.E., Reising, M.M., Leibenluft, E., et al., 2009. Amygdala activation during emotion processing of neutral faces in children with severe mood dysregulation versus ADHD or bipolar disorder. Am. J. Psychiatry 167 (1), 61–69.
Bunford, N., Kinney, K.L., Michael, J., Klumpp, H., 2017. Threat distractor and perceptual load modulate test-retest reliability of anterior cingulate cortex response. Prog. Neuro-Psychopharmacol. Biol. Psychiatry 77, 120–127.
Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C.R., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. NeuroImage 45 (3), 758–768. http://dx.doi.org/10.1016/j.neuroimage.2008.12.035.
Chen, G., Saad, Z.S., Britton, J.C., Pine, D.S., Cox, R.W., 2013. Linear mixed-effects modeling approach to FMRI group analysis. NeuroImage 73, 176–190.
Chen, G., Taylor, P.A., Haller, S.P., Kircanski, K., Stoddard, J., Pine, D.S., Cox, R.W., et al., 2017. Intraclass correlation: improved modeling approaches and applications for neuroimaging. Hum. Brain. Mapp. 39 (3), 1187–1206.
Collins, F.S., Tabak, A.L., 2014. NIH plans to enhance reproducibility. Nature 505 (7485), 612–613.
Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29 (3), 162–173.
Cox, R.W., Chen, G., Glen, D.R., Reynolds, R.C., Taylor, P.A., 2017. FMRI clustering in AFNI: false positive rates redux. Brain Connect. 7 (3), 152–171.
Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., Liu, J., 2013. A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. Psychometrika 78 (4), 685–709.
Ekman, P., Friesen, W.V., 1976. Pictures of Facial Affect. Consulting Psychologists Press, Palo Alto, CA.
Etkin, A., Egner, T., Kalisch, R., 2011. Emotional processing in anterior cingulate and medial prefrontal cortex. Trends Cogn. Sci. 15 (2), 85–93.
Fonville, L., Giampietro, V., Surguladze, S., Williams, S., Tchanturia, K., 2014. Increased BOLD signal in the fusiform gyrus during implicit emotion processing in anorexia nervosa. NeuroImage: Clin. 4, 266–273. http://dx.doi.org/10.1016/j.nicl.2013.12.002.
Garrett, A.S., Miklowitz, D.J., Howe, M.E., Singh, M.K., Acquaye, T.K., Hawkey, C.G., Chang, K.D., et al., 2015. Changes in brain activation following psychotherapy for youth with mood dysregulation at familial risk for bipolar disorder. Prog. Neuro-Psychopharmacol. Biol. Psychiatry 56, 215–220.
Gee, D.G., Humphreys, K.L., Flannery, J., Goff, B., Telzer, E.H., Shapiro, M., Tottenham, N., et al., 2013. A developmental shift from positive to negative connectivity in human amygdala–prefrontal circuitry. J. Neurosci. 33 (10), 4584–4593.
Hassel, S., Almeida, J.R., Frank, E., Versace, A., Nau, S.A., Klein, C.R., Phillips, M.L., et al., 2009. Prefrontal cortical and striatal activity to happy and fear faces in bipolar disorder is associated with comorbid substance abuse and eating disorder. J. Affect. Disord. 118 (1–3), 19–27. http://dx.doi.org/10.1016/j.jad.2009.01.021.
Johnstone, T., Somerville, L.H., Alexander, A.L., Oakes, T.R., Davidson, R.J., Kalin, N.H., Whalen, P.J., 2005. Stability of amygdala BOLD response to fearful faces over multiple scan sessions. NeuroImage 25 (4), 1112–1123. http://dx.doi.org/10.1016/j.neuroimage.2004.12.016.
Kober, H., Barrett, L.F., Joseph, J., Bliss-Moreau, E., Lindquist, K., Wager, T.D., 2008. Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies. Neuroimage 42 (2), 998–1031.
Kalmar, J.H., Wang, F., Chepenik, L.G., Womer, F.Y., Jones, M.M., Pittman, B., Blumberg, H.P., et al., 2009. Relation between amygdala structure and function in adolescents with bipolar disorder. J. Am. Acad. Child. Adolesc. Psychiatry 48 (6), 636–642. http://dx.doi.org/10.1097/CHI.0b013e19f6fbc.
Kaufman, J., Birmaher, B., Brent, D., Rao, U.M.A., Flynn, C., Moreci, P., Ryan, N., et al., 1997. Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADS-PL): initial reliability and validity data. J. the Am. Acad. Child. Adolesc. Psychiatry 36 (7), 980–988.
Kim, P., Thomas, L.A., Rosen, B.H., Moscicki, A.M., Brotman, M.A., Zarate Jr, C.A., et al., 2012. Differing amygdala responses to facial expressions in children and adults with bipolar disorder. Am. J. Psychiatry 169 (6), 642–649.
Lawrence, N.S., Williams, A.M., Surguladze, S., Giampietro, V., Brammer, M.J., Andrew, C., Phillips, M.L., et al., 2004. Subcortical and ventral prefrontal cortical neural responses to facial expressions distinguish patients with bipolar disorder and major depression. Biol. Psychiatry 55 (6), 578–587. http://dx.doi.org/10.1016/j.biopsych.2003.11.017.
Manuck, S.B., Brown, S.M., Forbes, E.E., Hariri, A.R., 2007. Temporal stability of

individual differences in amygdala reactivity. Am. J. Psychiatry 164 (10), 1613–1614.
Marusak, H.A., Zundel, C.G., Brown, S., Rabinak, C.A., Thomason, M.E., 2017. Convergent behavioral and corticolimbic connectivity evidence of a negativity bias in children and adolescents. Soc. Cogn. Affect. Neurosci. 12 (4), 517–525.
McLaren, D.G., Ries, M.L., Xu, G., Johnson, S.C., 2012. A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. Neuroimage 61 (4), 1277–1286.
Nord, C.L., Gray, A., Charpentier, C.J., Robinson, O.J., Roiser, J.P., 2017. Unreliability of putative fMRI biomarkers during emotional face processing. NeuroImage 156, 119–127.
Plichta, M.M., Schwarz, A.J., Grimm, O., Morgen, K., Mier, D., Haddad, L., Meyer-Lindenberg, A., et al., 2012. Test–retest reliability of evoked BOLD signals from a cognitive–emotive fMRI test battery. NeuroImage 60 (3), 1746–1758. http://dx.doi.org/10.1016/j.neuroimage.2012.01.129.
Rosenfeld, E.S., Pearlson, G.D., Sweeney, J.A., Tamminga, C.A., Keshavan, M.S., Nonterah, C., Stevens, M.C., 2014. Prolonged hemodynamic response during incidental facial emotion processing in inter-episode bipolar I disorder. Brain Imaging Behav. 8 (1), 73–86. http://dx.doi.org/10.1007/s11682-013-9246-z.
Sauder, C.L., Hajcak, G., Angstadt, M., Phan, K.L., 2013. Test-retest reliability of amygdala response to emotional faces. Psychophysiology 50 (11), 1147–1156. http://dx.doi.org/10.1111/psyp.12129.
Ségonne, F., Dale, A., Busa, E., Glessner, M., Salat, D., Hahn, H., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. NeuroImage 22 (3), 1060–1075.
Shah, M.P., Wang, F., Kalmar, J.H., Chepenik, L.G., Tie, K., Pittman, B., Blumberg, H.P., et al., 2008. Role of variation in the serotonin transporter protein Gene (SLC6A4) in trait disturbances in the ventral anterior cingulate in bipolar disorder. Neuropsychopharmacology 34 (5), 1301–1310.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 (2), 420.
Somerville, L.H., Kim, H., Johnstone, T., Alexander, A.L., Whalen, P.J., 2004. Human amygdala responses during presentation of happy and neutral faces: correlations with state anxiety. Biol. Psychiatry 55 (9), 897–903.
Stoddard, J., Tseng, W., Kim, P., et al., 2017. Association of irritability and anxiety with the neural mechanisms of implicit face emotion processing in youths with psychopathology. JAMA Psychiatry 74 (1). http://dx.doi.org/10.1001/jamapsychiatry.2016.3282.
Surguladze, S., Brammer, M.J., Keedwell, P., Giampietro, V., Young, A.W., Travis, M.J., Phillips, M.L., et al., 2005. A differential pattern of neural response toward sad versus happy facial expressions in major depressive disorder. Biol. Psychiatry 57 (3), 201–209. http://dx.doi.org/10.1016/j.biopsych.2004.10.028.
Surguladze, S.A., Marshall, N., Schulze, K., Hall, M.H., Walshe, M., Bramon, E., McDonald, C., et al., 2010. Exaggerated neural response to emotional faces in patients with bipolar disorder and their first-degree relatives. NeuroImage 53 (1), 58–64. http://dx.doi.org/10.1016/j.neuroimage.2010.05.069.
Thomas, L.A., De Bellis, M.D., Graham, R., LaBar, K.S., 2007. Development of emotional facial recognition in late childhood and adolescence. Dev. Sci. 10 (5), 547–558.
Thomas, L.A., Kim, P., Bones, B.L., Hinton, K.E., Milch, H.S., Reynolds, R.C., Leibenluft, E., et al., 2013. Elevated amygdala responses to emotional faces in youths with chronic irritability or bipolar disorder. NeuroImage: Clin. 2, 637–645. http://dx.doi.org/10.1016/j.nicl.2013.04.007.
Whalen, P.J., Shin, L.M., McInerney, S.C., Fischer, H., Wright, C.I., Rauch, S.L., 2001. A functional MRI study of human amygdala responses to facial expressions of fear versus anger. Emotion 1 (1), 70.
van den Bulk, B.G., Koolschijn, P.C.M.P., Meens, P.H.F., van Lang, N.D.J., van der Wee, N.J.A., Rombouts, S.A.R.B., Crone, E.A., et al., 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. http://dx.doi.org/10.1016/j.dcn.2012.09.005.
Wechsler, D., 1999. Wechsler Abbreviated Scale of Intelligence. Psychological Corporation.
White, L.K., Britton, J.C., Sequeira, S., Ronkin, E.G., Chen, G., Bar-Haim, Y., Pine, D.S., et al., 2016. Behavioral and neural stability of attention bias to threat in healthy adolescents. NeuroImage 136, 84–93. http://dx.doi.org/10.1016/j.neuroimage.2016.04.058.