NeuroImage 285 (2024) 120503

Contents lists available at ScienceDirect

NeuroImage

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

A precision neuroscience approach to estimating reliability of neural responses during emotion processing: Implications for task-fMRI

John C. Flournoy[a,*], Nessa V. Bryce[a], Meg J. Dennison[b], Alexandra M. Rodman[a], Elizabeth A. McNeilly[c], Lucy A. Lurie[d], Debbie Bitran[f], Azure Reid-Russell[a], Constanza M. Vidal Bustamante[a], Tara Madhyastha[e,g], Katie A. McLaughlin[a]

[a] Department of Psychology, Harvard University
[b] Phoenix Australia—Centre for Posttraumatic Mental Health, Department of Psychiatry, The University of Melbourne, Melbourne, VIC, Australia
[c] Department of Psychology, University of Oregon
[d] Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill
[e] Rescale
[f] Department of Psychology, University of Pittsburgh
[g] Integrated Brain Imaging Center, University of Washington

ARTICLE INFO

Keywords:
fMRI
Reliability
Neuroimaging
Longitudinal
Individual differences
Translational neuroscience
Clinical neuroscience

ABSTRACT

Recent work demonstrating low test-retest reliability of neural activation during fMRI tasks raises questions about the utility of task-based fMRI for the study of individual variation in brain function. Two possible sources of the instability in task-based BOLD signal over time are noise or measurement error in the instrument, and meaningful variation across time within-individuals in the construct itself—brain activation elicited during fMRI tasks. Examining the contribution of these two sources of test-retest unreliability in task-evoked brain activity has far-reaching implications for cognitive neuroscience. If test-retest reliability largely reflects measurement error, it suggests that task-based fMRI has little utility in the study of either inter- or intra-individual differences. On the other hand, if task-evoked BOLD signal varies meaningfully over time, it would suggest that this tool may yet be well suited to studying intraindividual variation. We parse these sources of variance in BOLD signal in response to emotional cues over time and within-individuals in a longitudinal sample with 10 monthly fMRI scans. Test-retest reliability was low, reflecting a lack of stability in between-person differences across scans. In contrast, within-person, within-session internal consistency of the BOLD signal was higher, and within-person fluctuations across sessions explained almost half the variance in voxel-level neural responses. Additionally, monthly fluctuations in neural response to emotional cues were associated with intraindividual variation in mood, sleep, and exposure to stressors. Rather than reflecting trait-like differences across people, neural responses to emotional cues may be more reflective of intraindividual variation over time. These patterns suggest that task-based fMRI may be able to contribute to the study of individual variation in brain function if more attention is given to within-individual variation approaches, psychometrics—beginning with improving reliability beyond the modest estimates observed here, and the validity of task fMRI beyond the suggestive associations reported here.

1. Introduction

Functional MRI (fMRI) was developed as a tool to investigate properties of brain function in humans. The classic approach to doing so involves contrasting the blood-oxygen-level-dependent (BOLD) signal while participants are engaged in a task designed to elicit a particular cognitive or affective state with BOLD signal during a relevant control condition (Huettel et al., 2004; Poldrack et al., 2011). This approach has stimulated substantial knowledge about the functional properties of specific brain regions (Garrison et al., 2013; Sergerie et al., 2008; Shenhav et al., 2013).

More recently, fMRI data have been used to examine differences in brain function across individuals. A recent meta-analysis and analysis of two large cohorts (Elliott et al., 2020) calls into question the validity of using task-based fMRI to study these types of between-person individual-differences in brain function, demonstrating low test-retest

[*] Corresponding author at: Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138, USA.
E-mail address: john_flournoy@g.harvard.edu (J.C. Flournoy).

https://doi.org/10.1016/j.neuroimage.2023.120503
Received 16 June 2023; Received in revised form 13 December 2023; Accepted 15 December 2023
Available online 22 December 2023

reliability for many common fMRI tasks. These findings suggest that the stability of task-related BOLD signal over time is relatively poor, which undermines the utility of fMRI as a brain-based biomarker (Elliott et al., 2020). Test-retest reliability was particularly poor for amygdala activation in tasks involving emotion processing, which have frequently been used to study individual differences in brain function. These types of tasks have been used frequently to investigate whether brain function varies as a function of mental health symptoms (Etkin and Wager, 2007; Groenewold et al., 2013), personality traits (Canli et al., 2001; Gray et al., 2005), or environmental experiences, such as early-life adversity (McLaughlin et al., 2019; Tottenham et al., 2011). Here, we investigate an alternative interpretation of this finding—that the BOLD signal measured in task-based fMRI is reliable, but exhibits high within-person variability over time.

The relatively low intra-class correlation coefficients (ICCs) observed in the recent meta-analysis (Elliott et al., 2020) mean that a participant who exhibits high activation in a particular region in response to a task at one time is not more likely to exhibit high activation in that same region when tested at a later time on the same task, relative to others in a sample. One possible source contributing to this variability in task-based BOLD signal over time is noise or measurement error in the instrument (Chen et al., 2018; Gonzalez-Castillo et al., 2017). In addition, the construct itself—brain activation elicited during fMRI tasks—may vary meaningfully across time within-individuals. In other words, variability in neural activation to tasks may be more state-like than trait-like. Indeed, influential recent work suggests that a wide range of psychological (e.g., affect) and physiological (e.g., heart rate) constructs exhibit greater variability within individuals than across individuals (Fisher et al., 2018). Variation in neural response to fMRI tasks may in part reflect high temporal variability (and so low test-retest reliability) in the construct itself, as well as measurement error. Examining the contribution of these two sources of test-retest unreliability in task-evoked brain activity has far-reaching implications for cognitive neuroscience. If test-retest reliability largely reflects measurement error, it suggests that task-based fMRI has little utility in the study of either inter- or intra-individual differences. On the other hand, if task-evoked BOLD signal as a construct varies meaningfully over time, it would suggest that this tool may still be well suited to studying intraindividual (i.e., within-individual) variation.

We use data from a unique longitudinal study involving monthly fMRI scans of an emotion processing task on the same individuals over the course of one year to characterize the degree of inter- and intra-individual variability in neural responses to affectively-salient cues and the reliability of these responses over time and the internal consistency of the BOLD signal within individuals. This intensive longitudinal approach aligns with the emerging field of precision neuroscience, which focuses on repeated sampling of fMRI data from the same individuals over time to examine patterns of stability and variability in neural function (Braga and Buckner, 2017; Gordon et al., 2017; Gratton et al., 2020; Laumann et al., 2015; Poldrack, 2017). Early precision neuroscience studies revealed dynamic fluctuations in brain function within individuals in networks previously thought to be stable based on between-participant designs (Poldrack et al., 2015). Influential recent work recommends within-participant longitudinal designs as a more powerful strategy for examining brain-behavior associations than cross-sectional brain-wide association studies (Marek et al., 2022). Neural responses during emotion processing may be particularly likely to vary within-individuals over time, given the high within-individual variation of affect (Rocke et al., 2009). Meta-analysis of these types of emotion processing tasks consistently reveal activation in amygdala as well as widespread cortical recruitment across regions in the salience network, such as anterior insula; the default network, including precuneus, posterior cingulate, medial prefrontal cortex, and middle temporal gryus; the frontoparietal network, including inferior, middle, and superior frontal gyrus; as well as the fusiform and other cortical regions in the ventral visual stream (Fusar-Poli et al., 2009; Sabatinelli et al., 2011). Despite these robust patterns at the group-level, the stability of neural responses in affective processing tasks over time was low in the recent meta-analysis (Elliott et al., 2020). To evaluate the degree to which this variability reflects meaningful within-individual fluctuations in neural responses versus measurement error, we estimate the internal consistency of the BOLD signal across the brain during an emotion processing task, within individuals, at each session. In doing so, we aim to contribute to the emerging debate about the reliability and utility of task-based fMRI for studying individual variation (Elliott et al., 2020; Kragel et al., 2021).

Characterizing the reliability of task-related BOLD signal is important because it puts an upper bound on our ability to detect valid associations with other measurements. In fact, an indirect indication of the reliability of a signal is the ability to detect expected associations with other constructs of interest. We are careful to note that there is not a one-to-one correspondence between reliability and the detection of associations. While higher reliability increases the ability to detect associations, larger effect sizes do the same; and while some reliability is necessary for detection, high reliability does not allow one to detect associations that are not present. However, detection of true associations suggests that reliability is high enough for a given effect size, with the important caveat that one does not know whether the detected effect is true or not. Moreover, as has been written about extensively elsewhere (Pashler and Harris, 2012), low reliability resulting in low statistical power increases the probability that any given result is a false positive, and with a publication filter on significant effects, inflates the proportion of false positives in the literature.

As such, we further investigate reliability by leveraging our dense sampling approach to determine whether we can predict variability in task-related BOLD response over time, within individuals. Specifically, we evaluate whether monthly fluctuations in mood, sleep quantity, and exposure to stressful life events (SLEs) are associated with changes in neural responses to affective cues over time, within individuals. These factors each fluctuate dynamically within individuals over time and have been associated with neural responses to aversive cues in between-person studies (Arnone et al., 2012; Goldstein-Piekarski et al., 2015; Larson and Ham, 1993; Mroczek and Almeida, 2004; Sliwinski et al., 2009; Swartz et al., 2015b, 2015a; Wang et al., 2006). These studies have demonstrated increases in activation of amygdala, anterior insula, and other regions of the salience network to aversive stimuli in individuals who have experienced high levels of SLEs (McLaughlin et al., 2019; Swartz et al., 2015b). Associations of sleep with neural responses during emotional processing are somewhat mixed and have largely focused on the amygdala. Greater sleep duration, particularly REM sleep, is associated with reduced amygdala reactivity to aversive stimuli (van der Helm et al., 2011; Wassing et al., 2019), whereas sleep deprivation predicts elevated amygdala reactivity (Yoo et al., 2007). A recent analysis from the UK Biobank observed the opposite pattern, however, with habitual short sleep associated with decreased amygdala reactivity (Schiel et al., 2022). Studies linking mood to neural responses to emotion processing have also observed that higher levels of negative affect are related to increased amygdala reactivity and decreased prefrontal recruitment to aversive stimuli (Bastiaansen et al., 2018; Forbes et al., 2011). However, influential recent work has shown that between-person associations often do not align well with within-individual associations of the same constructs for a range of psychological and physiological variables (Fisher et al., 2018). For this reason, we did not necessarily expect to find the neural regions that have been observed in between-person studies of mood, sleep, and SLEs to emerge in our within-person analyses. We are unaware of prior work examining whether fluctuations in mood, sleep, or SLEs are related to within-individual variation in neural responses to affective stimuli.

2. Method

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, John C. Flournoy (john_flournoy@g.harvard.edu; jcflournoyphd@pm.me).

2.1. Data and code availability

Code and data repository: https://osf.io/zy92w/.

- De-identified human standardized fMRI data have been deposited at repositories listed above. They are publicly available as of the date of publication.
- All original code has been deposited at repositories listed above and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request. Please let the authors know immediately if any resources are missing from the public repositories so they can be updated.

2.2. Sample

The study was designed to examine within-individual variation in neural responses to affective stimuli and associations of that variability with other key constructs that fluctuate over time within-individuals. A sample of 30 female adolescents aged 15–17 participated in a year-long longitudinal study that included 12 in-lab assessments conducted each month (355 monthly assessments) with neuroimaging completed at 10 of the 12 monthly visits, excluding the baseline and final visit (292 scans). Given our within-individual approach, we aimed to limit between-person variability in sex and age, focusing on adolescent females given the high levels of interpersonal stressors and stress vulnerability in this group (Hankin et al., 1998; Lewinsohn et al., 1998; Rudolph and Hammen, 1999). Participants were recruited from schools, libraries, public transportation, and other public spaces in the general community in Seattle, WA between April 2016 and April 2018. Inclusion criteria included female sex, aged 15–17 years at study onset, possession of a smart phone with a data plan, and English fluency.

Participants were excluded based on the following criteria: IQ < 80, active substance dependence, psychosis, presence of pervasive developmental disorders (e.g., autism), MRI ineligibility (e.g. metal implants), psychotropic medication use, active safety concerns, and inability to commit to the year-long study procedure.

Twenty-two participants identified as White (73%), 4 as Asian (13%), 2 as Black (7%), and 2 as mixed race (7%). Participants' income-to-needs ratios were computed based on their parents' report of total combined household income and household size. Four participants were in families with income below the poverty line (i.e., income-to-needs ratio below 1; 13%), 12 participants between 1 and 3 (30%), and 13 participants between 3 and 10 (33%). One participant did not provide income information. All study procedures were approved by the Institutional Review Board at the University of Washington. Written informed consent was obtained from legal guardians and adolescents provided written assent. Participants were paid increasing amounts of money for each monthly visit, for a total of $905 in possible earnings (Table S1).

2.3. Emotional processing task

Participants completed an emotional processing task involving passive viewing of emotional faces, including fearful, happy, neutral, and scrambled faces. We focus on the contrast of fearful > neutral faces, to examine variation in neural responses to aversive stimuli. This is a widely used contrast in affective neuroscience thought to capture neural responses to the presence of a potential threat in the environment. Indeed, similar tasks assessing neural responses to aversive stimuli have frequently been used in studies aimed at identifying brain-based biomarkers associated with stress, psychopathology, and numerous other between-person characteristics (Arnone et al., 2012; McCrory et al., 2011; Monk et al., 2008; Swartz et al., 2015b, 2015a; Thomas et al., 2001; Tottenham et al., 2011).

The task was completed across one run that included twelve 18-second blocks (three blocks each of fearful, happy, neutral, and scrambled faces). Blocks were displayed in a pseudo-random order that ensured that no block type was displayed twice in a row. ITI blocks were interleaved between blocks of faces. During each block, 36 faces of different actors expressing the same emotion were displayed for 300 ms each, with a space of 200 ms following each face, based on prior face processing tasks (Somerville et al., 2004). The total duration of the task was 4.5 min. The task was intentionally designed to be brief, given evidence that the amygdala, hippocampus, and other temporal cortex regions involved in emotional processing habituate rapidly to emotional faces (Breiter et al., 1996; Fischer et al., 2003). Similar brief tasks have been used to capture neural responses to affective stimuli in large-scale data collection efforts, such as the Human Connectome Project and the UK Biobank (Barch et al., 2013; Miller et al., 2016; Somerville et al., 2018).

Participants were asked to respond to prompts unrelated to the faces with a button press during the task to ensure they were paying attention. Specifically, an image (e.g., a scene or object) was displayed at one point in the run. A second image was subsequently presented and participants had to indicate whether the image was the same or different. Otherwise, participants were only asked to keep their eyes open and view the faces. Faces were drawn from the NimStim stimulus set (Tottenham et al., 2009). The "calm" faces from this dataset were used as neutral expressions, as these expressions are potentially less emotionally evocative than neutral faces, which are perceived as negatively-valenced (Tottenham et al., 2013). The scrambled faces consisted of the images of neutral faces with the pixels scrambled so as to resemble random static. Task-related functional activation for this contrast, averaged across all participants and months, controlling for the effect of time (centered at the first month), is depicted in Fig. S1.

2.4. Image acquisition and pre-processing

Neuroimaging data were acquired using a Phillips Achieva 3T scanner using a 32-channel head coil at the University of Washington's Integrated Brain Imaging Center. Anatomical scans (T1-weighted MPRAGE volumes; TR=2530ms, TE=3.5ms, flip angle=7°, FOV=256 × 256, 176 slices, in-plane voxel size=1mm3) were acquired for coregistration with functional magnetic resonance imaging (fMRI). Blood oxygenation level dependent (BOLD signal during functional runs was acquired using a gradient-echo T2*-weighted echo planar imaging (EPI) sequence. 37 3mm thick axial slices were acquired sequentially and parallel to the AC-PC line (TR=2s, TE=25ms, flip angle=79°, Interslice gap=.6mm, FOV=224 × 224 × 132.6, matrix size=76 × 74). Prior to each scan, four images were acquired and discarded to allow longitudinal magnetization to reach equilibrium.

Preprocessing was performed using FSL (Smith et al., 2004) and AFNI (Cox, 1996; Cox and Hyde, 1997), with the following steps applied to functional images: (a) motion and slice timing correction in FSL; (b) skull-stripping using AFNI's 3dSkullStrip; (c) despiking using AFNI's 3dDespike tool; and (d) for the purposes of whole-brain analysis, but not reliability and internal consistency analyses, smoothing with a 6mm full-width half-max kernel using SUSAN in FSL. Nuisance regressors entered into person-level models consisted of 6 rigid-body motion regressors as well as time-series extracted from white matter and ventricles entered to control for physiological noise (Behzadi et al., 2007). Outlier volumes in which framewise displacement exceeded 1mm, the derivative of variance in BOLD signal across the brain (DVARS) exceeded the upper fence, or signal intensity was more than 3 SD from the mean were excluded by regressing these volumes out of person-level models. First-level models were estimated on these preprocessed BOLD images. The resulting contrast images were registered first to a study-specific template (Ghosh et al., 2010; Huang et al., 2010; "Python Client - TemplateFlow," n.d.), and then to the standard space of the

Montreal Neurological Institute (MNI) template at 2mm resolution. Anatomical co-registration of the functional data from each monthly assessment with each participant's T1-weighted image from that same monthly assessment and normalization were performed using Advanced Normalization Tools (ANTs). All transforms were concatenated so only a single interpolation would be performed.

2.5. *Quantification and statistical analysis*

2.5.1. *Reliability analyses*

The term "reliability" can be used to refer to many kinds of consistencies between measurements, all of which have the shared goal of capturing how similar another measurement is expected to be using the same instrument (see Revelle and Condon, 2019 for a thorough review). Here, we focus on test-retest reliability and internal consistency as quantifications of two distinct kinds of consistency or reliability in the BOLD signal (though each is commonly used to index an instrument's signal-to-noise ratio). We describe our approach to estimating each of these in detail below.

Anatomical Regions. Analyzing parcellated brain data can aid complete reporting of effect sizes across the whole brain, and increase statistical power (Cosme et al., 2022; Flournoy et al., 2020). For this reason, reliability and internal consistency were assessed using the Schaefer 400 cortical parcellation scheme developed using both task-based and resting-state fMRI methods (Schaefer et al., 2018), as well as 14 anatomically-defined subcortical areas in the Harvard/Oxford subcortical atlas (brainstem; right and left accumbens, amygdala, hippocampus, caudate, pallidum, putamen, thalamus) and 24 size-matched control regions defined as spheres in right and left cerebral white matter and lateral ventricle from the Harvard/Oxford subcortical atlas. We included these control regions as a baseline comparison because we expect BOLD signal to have lower test-retest reliability and internal consistency in these regions than in cortical and subcortical regions. Each of these regions-of-interest (ROIs) were registered to each participant's monthly T1 image using ANTs, as described above.

To create size-matched control ROIs in left and right white-matter and ventricles, we used a random subsample of 10 cortical parcels and all subcortical regions ($N = 14$) to define the size of ROI for size-matching. Our procedure was as follows.

1. Compute the volume, in voxels, of the target subcortical region or cortical parcel.
2. Randomly select a voxel coordinate from the control region (left or right white matter or ventricle).
3. Construct a sphere with a radius such that the volume of the sphere is equal to the volume of the target region or parcel.
4. Ensure that the sphere falls within the control region. If not, repeat steps 1-3.
5. Compute the relevant statistic using voxels within the defined sphere.

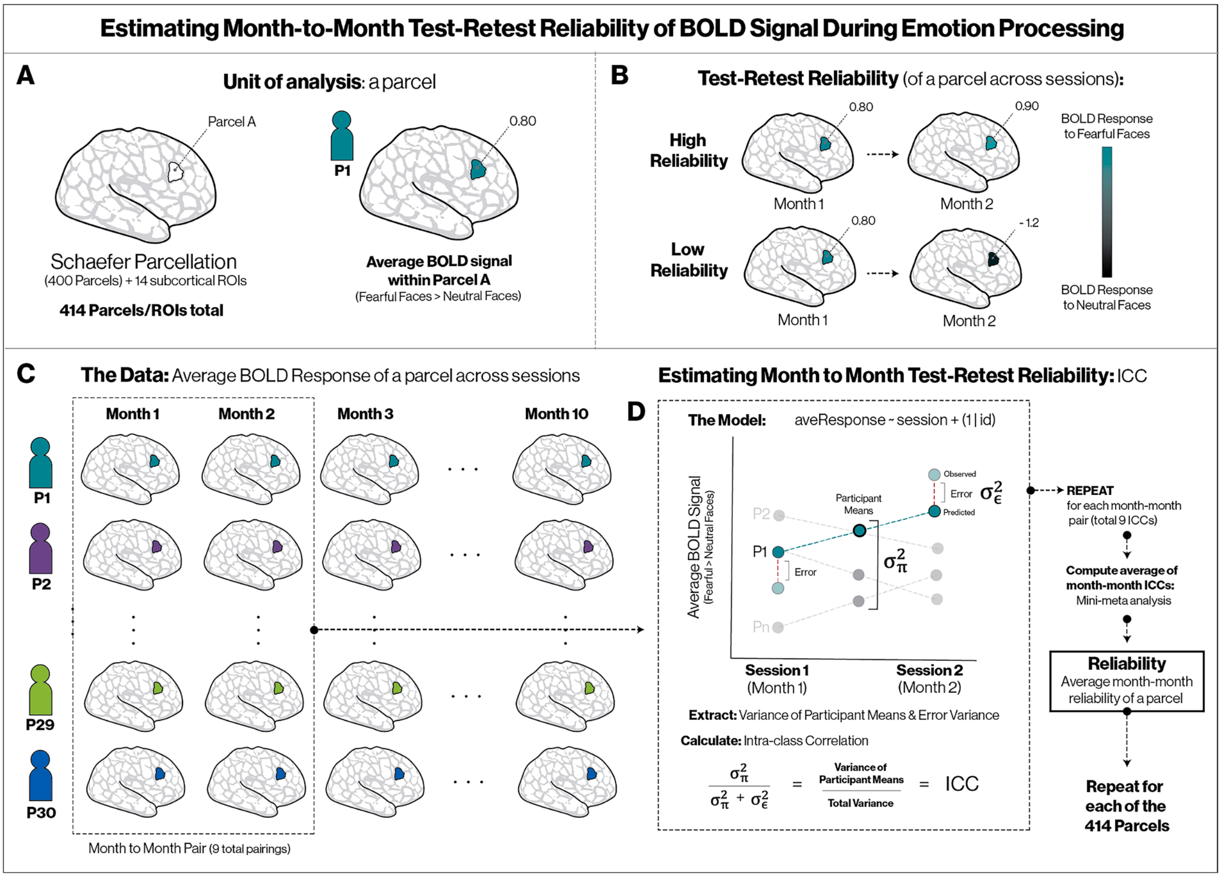
Estimating Month-to-Month Test-Retest Reliability of BOLD Signal During Emotion Processing
A
Unit of analysis: a parcel
Parcel A
P1
0.80
Schaefer Parcellation
(400 Parcels) + 14 subcortical ROIs
414 Parcels/ROIs total
Average BOLD signal within Parcel A
(Fearful Faces > Neutral Faces)
B
Test-Retest Reliability (of a parcel across sessions):
High Reliability
0.80
0.90
Month 1
Month 2
Low Reliability
0.80
-1.2
Month 1
Month 2
BOLD Response to Fearful Faces
BOLD Response to Neutral Faces
C
The Data: Average BOLD Response of a parcel across sessions
Month 1
Month 2
Month 3
Month 10
P1
P2
P29
P30
Month to Month Pair (9 total pairings)
Estimating Month to Month Test-Retest Reliability: ICC
D
The Model: aveResponse ~ session + (1 | id)
Average BOLD Signal
(Fearful > Neutral Faces)
P2
P1
Pn
Participant Means
Observed
Error
Predicted
σ^2_ϵ
σ^2_π
Error
Session 1
(Month 1)
Session 2
(Month 2)
Extract: Variance of Participant Means & Error Variance
Calculate: Intra-class Correlation
$\frac{\sigma^2_\pi}{\sigma^2_\pi + \sigma^2_\epsilon} = \frac{\text{Variance of Participant Means}}{\text{Total Variance}} = \text{ICC}$
REPEAT for each month-month pair (total 9 ICCs)
Compute average of month-month ICCs: Mini-meta analysis
Reliability
Average month-month reliability of a parcel
Repeat for each of the 414 Parcels

Fig. 1. Methods used to estimate test-retest reliability. A: The unit of analysis is the mean of voxels in a cortical parcel or subcortical region of interest (ROI) for the contrast for Fear > Neutral measured for each participant at each month. B: Test-retest reliability (i.e., temporal stability) is operationalized as the similarity (in rank order of participants) of the BOLD signal in each parcel from month to month. C: The data are observations for a particular parcel from each participant, for each month; we estimate an ICC for each pair of adjacent months for that parcel. This yields N x S rows of data for each parcel, where N is the sample size and S is the number of sessions (months). D: To estimate test-retest reliability, we compute an ICC that decomposes the total variance into the variance in participant means across months and error variance, after conditioning on the group means for each session. The ICC is then computed as the variance due to participant means over the total variance. We do this for each pair of adjacent months and then compute an average ICC across all pairs using meta-analysis. We compute this overall ICC as the measure of test-retest reliability for each of the 414 parcels and subcortical ROIs, and 5 regions of no-interest (see Methods for details).

Test-Retest Reliability. Test-retest reliability—a measure of temporal stability across participants in a sample (Chen et al., 2021, 2018; Koo and Li, 2016; Shrout and Fleiss, 1979)—was estimated by computing the ICC for BOLD signal in the Fear > Neutral contrast for each participant, at each session, in each cortical parcel and subcortical region (see Fig. 1). A Bayesian multilevel model (i.e., hierarchical linear model) was fit for each pair of adjacent sessions (e.g., sessions 1 and 2, sessions 7 and 8); estimating a single ICC across all 10 sessions for each participant produced consistently lower ICC estimates than this approach (see Fig. S2). Using *brms* (version 2.15.0; Bürkner, 2018, 2017) in R (v4.04; R Core Team, 2021) we fit the model

$$y_{ij} = \beta_0 + \beta_i + \pi_j + \epsilon_{ij}$$

where β_0 is the overall mean (i.e., the grand mean across all participants), β_i is the fixed effect of session *i*, and π_j is the random intercept of participant *j*. The ICC was then calculated as

$$ICC_{pair} = \frac{\sigma_\pi^2}{\sigma_\pi^2 + \sigma_\epsilon^2}$$

where σ_π^2 is the variance of π (i.e., random intercept variance) and σ_ε^2 is the residual (i.e., error) variance. ICC_{pair} corresponds to the ICC(3,1), also referred to as consistent agreement, (Chen et al., 2021, 2018; Koo and Li, 2016; Shrout and Fleiss, 1979) and reflects the proportion of variance due to participant means across sessions, and is also the expected correlation between observations across sessions from the same participant (Chen et al., 2018). Bayesian estimation was chosen primarily because it allows straightforward computation of credible intervals of the quantities of interest. The resulting posterior distributions were logit-transformed to rescale them from [0,1] to (-∞, ∞) and the medians and standard deviations were meta-analyzed using brms to obtain a single estimate of individual consistency between temporally adjacent sessions. The meta-analytic model was

$$\begin{aligned} \text{ICC}_{pair} &\sim \text{Normal}(\text{ICC}_{true}, \text{ICC}_{SE}) \\ \text{ICC}_{true} &\sim \text{Normal}(\mu, \sigma) \end{aligned}$$

where ICC_{pair} is the set of observed pairwise ICC estimates, ICC_{true} is the latent true ICC, ICC_{SE} is the set of uncertainties (standard deviations) in the observed ICC_{pair} estimates, and μ and σ are the mean and standard deviation, respectively, of ICC_{true}. The resulting posterior of parameter μ was logit^{-1}-transformed to rescale back to [0,1] and the resulting median and 95% credible interval were interpreted as estimates of test-retest reliability of neural responses for each region across all individuals in the sample.

This approach to estimating reliability evaluates the stability of neural responses in each region within participants over time by evaluating how much the data from each participant deviates from their person-level mean across sessions. Higher estimates reflect higher test-retest reliability in BOLD signal across participants (i.e., a high ICC implies that a participant with high BOLD signal in a particular region in a particular session, relative to others in the sample, is also likely to have high BOLD signal in that region in other sessions).

Internal Consistency. Low test-retest reliability is partly a function of the sources of variance that contribute to the magnitude of the error variance term, σ_ε^2, in the denominator. This term captures all unmodeled variance, and can reflect a mixture of high measurement error in measuring BOLD signal and high within-individual variability in responses over time. The internal consistency of multiple indicators of the same construct, based on generalizability theory (Bonito et al., 2012), is an alternative to the test-retest approach to quantifying instrument consistency that can disaggregate these sources of error. Reliable indicators that exhibit high within-individual variability (i.e., that fluctuate together over time), would produce high internal consistency, but low test-retest correlations. In these data, we can disaggregate variance not only into temporal stability across participants (reflecting test-retest correlations), but also into consistency across voxels at each session (reflecting internal consistency). This is one way to begin to approach quantifying the extent to which low test-retest reliability is a result of measurement error versus fluctuations of an internally consistent signal over time (Fig. 2). This approach is similar to evaluating the internal consistency of items on a scale by computing Cronbach's alpha, which is based on the average inter-item correlations (Fig. S3). Here, instead of items on a scale we evaluate the stability of BOLD signal across voxels within distinct anatomical regions. As we would do with items on a self-report scale, we use voxel responses within person, within session, to estimate the reliability of the measurement instrument (in this case, parcels and subcortical regions). If a scale has good internal consistency, when the value on a single item increases for a participant at a particular administration, the values of other items on the scale should similarly increase. Equivalently, if the BOLD signal is measured consistently across voxels within individual, within session, when the contrast value in a voxel increases for a participant in a particular session, the value of other voxels in that parcel or subcortical region should similarly increase (Fig. S3).

Internal consistency was estimated for the same set of parcels and subcortical regions described above to determine the proportion of variance accounted for by consistency in BOLD signal across the voxels within each region for each participant at each session. This approach examines the degree to which BOLD signal for the voxels in a single parcel or subcortical region fluctuate consistently with one another within each session for each participant (i.e., participant-sessions). To do so, we fit a multilevel model using *brms* to data from each parcel across all sessions with the form

$$y_{ijk} = \beta_0 + \beta_i + \lambda_j + \pi_{jk} + \epsilon_{ijk},$$

where β_0 is the overall mean, β_i is the fixed effect of session *i*, λ_j is the random intercept of participant *j*, and π_{jk} is the random intercept for each participant *j*'s session *k*. From this we get estimates of the variance in each participant's mean across sessions (σ_λ^2) the variance due to means of voxels (as deviations from participant means) within each participant-session (σ_π^2), and error variance (σ_ε^2)—the variance in deviations of each voxel value from the value implied by participant means plus participant-session means. The proportion of variance due to the mean of each participant-session reflects internal consistency of voxels within a parcel, which we calculate as

$$\text{ICC}_{within} = \frac{\sigma_\pi^2}{\sigma_\lambda^2 + \sigma_\pi^2 + \sigma_\epsilon^2}.$$

The medians and 95% credible intervals of the resulting posterior distributions of ICC_{within} were interpreted as the voxel-to-voxel consistency of within-individual, within-session neural responses in each parcel.

This approach estimates how much the data from voxels within a specific region deviate from the mean from all voxels in that region, for that participant in that session; this estimate is conceptually similar to the internal consistency of items in a scale commonly used as an estimate of the reliability of self-report scales. Higher estimates reflect higher consistency in BOLD signal across voxels in a parcel for each participant, at each session (i.e., the BOLD signal varies in a consistent way for all voxels in a particular parcel or subcortical region for each participant, at each session). Note that unlike the above $\text{ICC}_{\text{within}}$ equation, the equation for multilevel internal consistency also divides the error variance term by the number of items. Given that the number of voxels in each of our parcels is large (median number of voxels per parcel = 301, IQR 221 – 389), we report the raw proportion of variance instead, which is a more conservative approach. Also note that this metric was evaluated using the unsmoothed data. Smoothing is a standard pre-processing step in task-based fMRI analysis, but will inflate estimates of the true variance ratio within a parcel because it eliminates some amount of the unshared (i.e., error) variance. In order to further ameliorate the

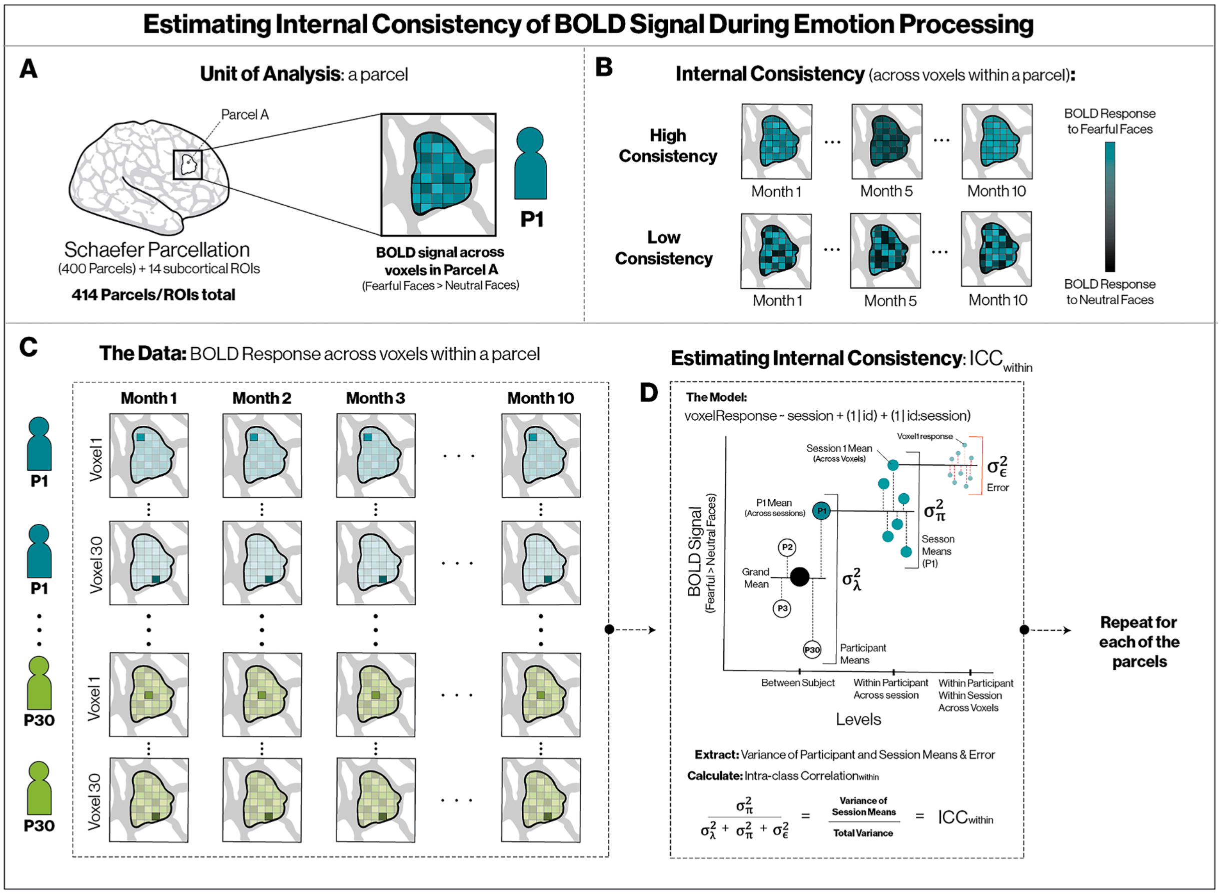
Estimating Internal Consistency of BOLD Signal During Emotion Processing
A
Unit of Analysis: a parcel
Parcel A
Schaefer Parcellation
(400 Parcels) + 14 subcortical ROIs
414 Parcels/ROIs total
BOLD signal across voxels in Parcel A
(Fearful Faces > Neutral Faces)
P1
B
Internal Consistency (across voxels within a parcel):
High Consistency
Low Consistency
Month 1
Month 5
Month 10
BOLD Response to Fearful Faces
BOLD Response to Neutral Faces
C
The Data: BOLD Response across voxels within a parcel
Month 1
Month 2
Month 3
Month 10
Voxel 1
Voxel 30
P1
P30
Estimating Internal Consistency: ICC within
D
The Model:
voxelResponse ~ session + (1 | id) + (1 | id:session)
Voxel1 response
Session 1 Mean (Across Voxels)
σ²ϵ
Error
P1 Mean (Across sessions)
σ²π
Sesson Means (P1)
Grand Mean
σ²λ
Participant Means
P1
P2
P3
P30
BOLD Signal (Fearful > Neutral Faces)
Between Subject
Within Participant Across session
Within Participant Within Session Across Voxels
Levels
Extract: Variance of Participant and Session Means & Error
Calculate: Intra-class Correlation within
σ²π / (σ²λ + σ²π + σ²ϵ) = Variance of Session Means / Total Variance = ICC within
Repeat for each of the parcels

Fig. 2. Methods used to estimate internal consistency. A: The unit of analysis is the voxel contrast parameter value within a cortical parcel or subcortical region of interest (ROI) from the contrast for Fear > Neutral measured for each participant at each month. B: Internal consistency is operationalized as the similarity (in rank order of voxels) of the signal across participant-sessions; high internal consistency means that voxels within a parcel for one participant-session tend to be similar to one another as compared to variation across participant-sessions. C: The data are observations for a particular voxel within a parcel from each participant, for each month; we estimate an ICC based on all voxels in that parcel. This yields N x S x V rows of data for each parcel, where N is the sample size, S is the number of sessions (months), and V is the number of voxels in the parcel. D: To estimate internal consistency, we compute an ICC that decomposes variance into the variance in participant means across all months, the mean for each session for each participant, and error variance, after conditioning on the group means for each session. The ICC is computed as the variance in participant-session means over the total variance. This is computed for each of the 414 parcels and subcortical ROIs, and 5 regions of no-interest (see Methods for details).

potential influence of smoothness we additionally computed ICCs for random subsets of 15 voxels from each parcel rather than from every voxel within a parcel.

Other approaches for estimating reliability of BOLD signal in task fMRI have been developed for event-related designs using individual trials as the unit of measurement (Chen et al., 2021). These methods cannot be applied here given the block design of our task and the strong habituation that occurs in both cortical and subcortical regions during emotion processing tasks (Breiter et al., 1996; Fischer et al., 2003).

Smoothness and Internal Consistency. Given that other properties of the BOLD signal aside from task-related neural responses may contribute to correlations of voxels that are spatially proximal, we consider these estimates as an upper bound of the amount of variance in BOLD signal over time that could reflect internally consistent within-individual variation. To probe whether these other properties of the BOLD signal are driving internal consistency estimates, we additionally assess the association between smoothness of the spatial signal and our measure of internal consistency. For each cortical parcel, subcortical ROI, and size-matched control region, for each participant, for each session, we estimated the smoothness of the first-level residuals using 3dFWHMx in AFNI (Cox, 1996; Cox and Hyde, 1997). We then regressed the estimates of internal consistency on the number of voxels in each parcel and the average smoothness of each parcel, each allowed to vary as a function of anatomy classified as cortex, subcortex, and control region. Each predictor was encoded using smooth functions (Wood, 2017), and the internal consistency outcome was modeled as distributed beta to account for the bounded values in [0, 1]. We compared this model to a constrained model including only the linear effect of parcel size and random intercept by anatomy (see supplemental text for more details). Model comparison employed efficient approximate leave-one-out cross-validation (Vehtari et al., 2018), which provides an expected log pointwise predictive density difference (ΔELPD) between models as well as standard errors of that difference. Models were considered non-equivalent if the absolute size of the ΔELPD exceeded two standard errors, and the simplest (i.e., constrained) model was retained in absence of evidence of non-equivalence.

2.5.2. Predictors of longitudinal within-individual variation in brain activity

Mood. Negative affect was assessed using the Positive and Negative Affect Scale (PANAS), a twenty-item measure assessing positive and negative affect (Watson et al., 1988). The general form of the PANAS was used, and participants were asked to indicate the extent to they felt this way over the past month. Participants respond to each affective state (e.g., excited, interested, nervous, hostile) on a 5-point Likert scale ranging from 1 (very slightly or not at all) to 5 (extremely). The PANAS has excellent internal consistency and has demonstrated convergent, discriminant, and predictive validity in a number of investigations (Watson et al., 1988; Watson and Walker, 1996). The PANAS was administered at each monthly visit to assess mood over the month since the previous visit.

Sleep. Daily sleep duration was assessed via an actigraphy wristband participants wore continuously for the duration of the study. The wristbands used accelerometer data collected in 1-min epochs and proprietary algorithms to detect sleep and awake states. These devices have been validated against polysomnography and EEG, with excellent sensitivity (i.e., ability to detect true sleep), and adequate specificity (i.e., ability to detect true wake; de Zambotti et al., 2016; Liang et al., 2018). We collected a total of 6824 daily sleep observations. We computed daily sleep duration in hours, aggregating over potentially multiple sleep events in the same day (e.g., naps, fragmented night sleep), and not including awake time between sleep events. Daily sleep duration was computed in hours each day for the 24-h periods from 7pm to 7pm, and then averaged across the two weeks occurring prior to each scan session. See supplement and Vidal Bustamante and colleagues (2020) for more information on the actigraphy devices and missing data.

Stressful Life Events (SLEs). SLEs occurring in the past month were assessed at each visit using the UCLA Life Stress Interview, a semi-structured interview designed to objectively measure the impact of life events (Hammen, 1988). The interview assesses acute life events/episodic stressors (e.g., failing a test, break-up of a romantic relationship) and chronic stress (e.g., ongoing conflict in the home, long-term medical issues). The interview has been extensively validated and adapted for use in adolescents (Daley et al., 1997; Dohrenwend, 2006; Dohrenwend and Shrout, 1985, 1985; Hammen, 1991). Structured prompts are used to query numerous domains of life (i.e., peers, parents, household/extended family, neighborhood, school, academic, health, finance, and discrimination). Each reported stressor is probed to determine timing, duration, severity, and coping resources available. Research personnel objectively coded the severity of each experience on a 9-point scale ranging from 1 (none) to 5 (extremely severe), including half-points. Following prior work, a total episodic stress score was computed by taking the sum of the severity scores of all reported events, which reflects both the number and severity of acute stressors (Hammen et al., 2000), hereafter referred to as SLEs. If the participant did not report any SLEs, they received a score of zero for that month. The severity of chronic stressors occurring in each domain were coded on the same scale. The chronic stress score for the domain where the participant was experiencing the highest amount of ongoing stress was used in analysis. The interview was administered at each monthly visit to assess SLEs and chronic stress occurring since the previous visit.

2.5.3. Modeling longitudinal within-individual variation in brain activity

To evaluate whether mood, sleep, and SLEs were associated with within-individual variation in neural response to aversive cues, we implemented voxel-wise and parcel-level multilevel models designed to disaggregate between- and within-person variation in longitudinal data. A frequentist power analysis from simulation indicates that we have 80% power to detect a standardized regression coefficient of .17 (see supplement for power across effect sizes and further details).

Neuropointillist. Traditional software packages for analyzing task-based fMRI data are limited in the types of statistical models that can be estimated to examine predictors of task-related activation and involve a number of meaningful limitations when examining longitudinal data—for a lengthy discussion of this issue, see Madhyastha and colleagues (2018). As a result, most longitudinal fMRI studies using complex longitudinal modeling approaches have extracted BOLD signal from ROIs and estimated models outside of fMRI software packages (Braams et al., 2015; Ordaz et al., 2013), limiting the analysis to specific ROIs rather than taking a whole-brain voxel-wise approach, which remains a common approach to analyzing fMRI data in cognitive neuroscience.

To address this issue, one of the authors of this work (TM) developed an R package ‘Neuropointillist’ that allows any model that can be specified in R to be estimated in fMRI data in each voxel of the brain (Madhyastha et al., 2018). Neuropointillist assembles longitudinal pre-processed and spatially normalized longitudinal fMRI data into a long-form dataset, where each row represents data from a particular voxel in a particular participant at a particular time. Neuropointillist accepts a model to be executed on each voxel in the dataset, written as a function called from R. The specified model is applied to every voxel, for every participant, and each measurement occasion. This affords complete flexibility to evaluate any statistical model of interest for voxel-wise fMRI analysis, including longitudinal models. The statistical parameter estimates obtained from first-level analyses can also be imported into traditional fMRI packages. See Fig. S4 for greater detail. Neuropointillist is freely available; for details and documentation see: http://github.com/IBIC/neuropointillist.

Person-level models. Person-level models were estimated in FSL. Task-related regressors were created by convolving a boxcar function of phase duration with the standard (double-gamma) hemodynamic response function for each condition of the task (fear, happy, neutral, scrambled). A general linear model was constructed for each participant.

Longitudinal Analysis, voxel-level. Individual-level estimates of BOLD activity were submitted to group-level random effects models using Neuropointillist (Madhyastha et al., 2018). Voxel-wise models were implemented in R (R Core Team, 2021) using the *nlme* package (Pinheiro et al., 2020) using restricted maximum likelihood (REML), with the intercept allowed to vary randomly across participants. This method is robust to bias when data are missing at random (Matta et al., 2018). We did not employ Bayesian estimation in this case due to its high computational cost. We first estimated unconditional models including a term only for time to examine linear changes in voxel-wise BOLD signal across the ten sessions. To dissociate between- and within-person effects of SLEs, we used within-individual centering (i.e., centering each participant’s monthly observations around their person-specific mean across the year-long study period) and between-participant centering at the year-level (i.e., centering each participant’s mean for the entire study period relative to the overall mean for the entire sample). This approach orthogonalizes variation in a given predictor into between- and within-person components (Enders and Tofighi, 2007), accounting for the dependent nature of the data both over time and within-participant, while controlling for trait-level characteristics of the predictor (i.e., average level of negative affect, sleep duration, or severity of SLEs across the entire year). We estimated five models predicting BOLD signal for the contrast of fearful > neutral faces using the following predictor, each decomposed into between- and within-individual components as described above: 1) negative affect; 2) positive affect; 3) sleep duration; 4) acute SLEs; and 5) chronic stress. This approach allowed us to examine variation in BOLD signal as a function of within-individual variation in each of these factors after controlling for average between-person differences in each.

Voxel-wise models included a main effect of time (coded as the number of months since the first study visit), and these within-person and between-person centered stress variables as fixed effects. We used the *clubSandwich* package (Pustejovsky, 2019) to compute cluster-robust standard errors in the presence of possible autocorrelation (Pustejovsky and Tipton, 2018), and apply the Satterthwaite correction to the degrees of freedom used to compute the *p*-value of the coefficient test. This *p*-value was converted to a *Z*-score and used as the test-statistic.

To correct for multiple comparisons in whole-brain analyses, family-wise error (FWE) rate was controlled at α=.05 for each model using Equitable Thresholding and Clustering (ETAC) cluster correction implemented in AFNI (Cox, 2019). The ETAC method allows detection of both small and large clusters by establishing multiple combined cluster-forming p-value/cluster-size thresholds that together control the FWE across permuted brain maps. For each model, 1000 permutations were generated. The resulting 1000 permuted z-score maps were then used to generate (using *3dXClustSim*) and apply (using *3dMultithresh*) the ETAC thresholds to the statistical parameter maps from the group-level analysis.

We generated 1000 permutations for each image using the following procedure. Specifically, first, permutations matrices were generated appropriately for nested data by shuffling observation indexes within participants (Winkler et al., 2015, 2014). We then implemented the procedure described by Freedman and Lane (Freedman and Lane, 1983) as follows:

1. regress the dependent variable (Y) on covariates (i.e., time, and group-centered mean scores of the dependent variable), saving the residuals and the predicted values of Y;
2. permute the residuals according to the ith row of the permutation matrix, then add the permuted residuals to the predicted Y values to produce Y*;
3. regress Y* on the within-person centered variable of interest, X, and covariates, and save the permutation test statistic for the association between X and Y*.

As in the group-level model, we used cluster-corrected standard errors to derive a *t*-statistic with Saterthwaite-approximated degrees of freedom. The *p*-value of this *t*-statistic was transformed to a *Z*-score, and saved as the permutation test statistic. The resulting 1000 permuted *Z*-score maps were then used to generate (using *3dXClustSim*) and apply (using *3dMultithresh*) the ETAC thresholds to the statistical parameter maps from the group-level analysis.

Significant clusters reveal regions of the brain where BOLD signal systematically increased or decreased during months when participants had greater negative affect, positive affect, sleep duration, or exposure to stress than was typical for them across the year.

Longitudinal Analysis, parcel-level. For each parcel in the Schaefer 400 cortical parcellation scheme (Schaefer et al., 2018), as well as the 18 anatomically-defined subcortical areas in the Harvard/Oxford, we extracted voxel-level estimates for each participant-session. We then estimated the same model as above using bayesian estimation to be consistent with the reliability and internal consistency analyses at the parcel level in R (R Core Team, 2021) with *brms* (version 2.15.0; Bürkner, 2018, 2017), which averages across the voxels in a parcel in a model-based way. We obtained a posterior probability distribution for the estimate of each parameter of interest. Using this posterior distribution, we threshold the image so that the combined probability of making an error in the sign (Gelman and Carlin, 2014) of any of the coefficients is constrained to be less than 5%. To implement this, we first compute the proportion of the posterior distribution that has the same sign as the median, ordering the parcels from the largest to smallest value. We then compute the cumulative product, which is the probability of not making a sign error. We decide to interpret and display the set of coefficients that are most likely in the correct direction and that maintain the probability of sign error at < 5%. We also present uncorrected analyses examining effects in left and right amygdala for each predictor because so much previous work has focused on this region. These results are presented in the supplement.

3. Results

3.1. Test-retest reliability

We first estimated ICCs across adjacent sessions to examine test-retest reliability (i.e., temporal stability) of neural activation in response to aversive cues (fearful > neutral faces) within 400 cortical parcels, 14 subcortical regions, and 4 control regions (see Fig. 1 and Methods for details). Higher ICCs reflect higher test-retest reliability in BOLD signal across participants (i.e., a participant with high BOLD signal in a particular region in a particular session is also likely to have high BOLD signal in that parcel on other sessions, relative to others in the sample).

ICCs for test-retest reliability were uniformly small in magnitude and close to zero: ranges and interquartile intervals (IQR) for median posterior ICCs were ICC = [.04, .15] (IQR .06-.08; *N* = 400) across parcels, ICC = [.05, .12] (IQR .07-.08; N = 14) across sub-cortical regions, and ICC = [.04, .12] (IQR .06-.08; *N* = 24) across size-matched control regions (see Fig. 3A). This pattern indicates that the test-retest reliability of BOLD signal in response to aversive stimuli (fearful > neutral faces) is uniformly low across the brain.

3.2. Internal consistency

Next, we estimated the internal consistency of parcels and subcortical regions across voxels (see Fig. 2 and Methods for details). We used the same cortical parcels, subcortical regions, and control regions used to compute test-retest reliability to evaluate the internal consistency of the signal. Higher ICCs reflect greater consistency in BOLD signal across voxels in a particular region for each participant, at each session (i.e., the BOLD signal varies in a consistent way for all voxels in a particular region for each participant, at each session). The degree of internal consistency would indicate whether signal changes across these cortical parcels and subcortical regions reflect coherent signal rather than random fluctuations, and so inform the degree to which measurement error and high within-individual variation contributes to poor test-retest reliability in neural responses to aversive cues (fearful > neutral faces).

ICCs were substantially higher in magnitude than for test-retest reliability: ranges and IQRs for median posterior ICCs were ICC = [.13, .70] (IQR .37-.51) across cortical parcels, ICC = [.22, .45] (IQR .26-.34) across sub-cortical regions of interest, and ICC = [.14, .57] (IQR .28-.41; *N* = 24) in size-matched control regions (see Fig. 3B). These patterns demonstrate substantially higher within-person, within-session internal consistency than test-retest reliability. Estimates in subsamples of 15 voxels within each cortical parcel or subcortical ROI correlated nearly perfectly (r = .96) with estimates from all voxels. We interpret these patterns to suggest that the low test-retest reliability across sessions reflects, in part, internally consistent intraindividual variability in neural responses to aversive cues over time.

As a point of reference, we additionally compute Cronbach's *alpha* as another commonly used metric of internal consistency. The primary difference between the ICC and and *alpha* is that *alpha* scales the error variance by the inverse of the number of items. As would be expected, this increases the estimate when using all voxels, α = [0.98, 1.00] (IQR 0.99-1.00), or even when a subset of 15 voxels are selected randomly from parcels (in order to keep scaling factor of the error variance reasonably small), with α = [0.65, 0.97] (IQR 0.90-0.94) for cortical parcels, and α = [0.81, 0.93] (IQR 0.84-0.88) for subcortical regions.

3.3. Smoothness and internal consistency

Given the spatial nature of the analysis, to determine how smoothness contributes to the internal consistency observed we conducted additional analyses to examine the association of smoothness with estimates of internal consistency. A model including smoothness did not fit significantly better than a model with only parcel size and anatomy, and in fact fit worse (ΔELPD = −31.2, SE = 23.9; see Methods for a description of the ELPD), indicating that BOLD signal smoothness was not associated with internal consistency estimates after controlling for parcel size; this is also visible as nearly flat trends in the effect of smoothness in the plots (Fig. 4A; see full model output in supplement, and Fig. S5 for zero-order association of all three predictor variables). The model-expected internal consistency (from the constrained model) for each anatomy type at the median parcel size (312 voxels) were: cortex, .44 (95% CI = [.43, .45]); subcortex, .36 (95% CI = [.31, .42]); and size-matched control, .35 (95% CI = [.30, .41]; Fig. 4B), suggesting greater internal consistency in cortical parcels.

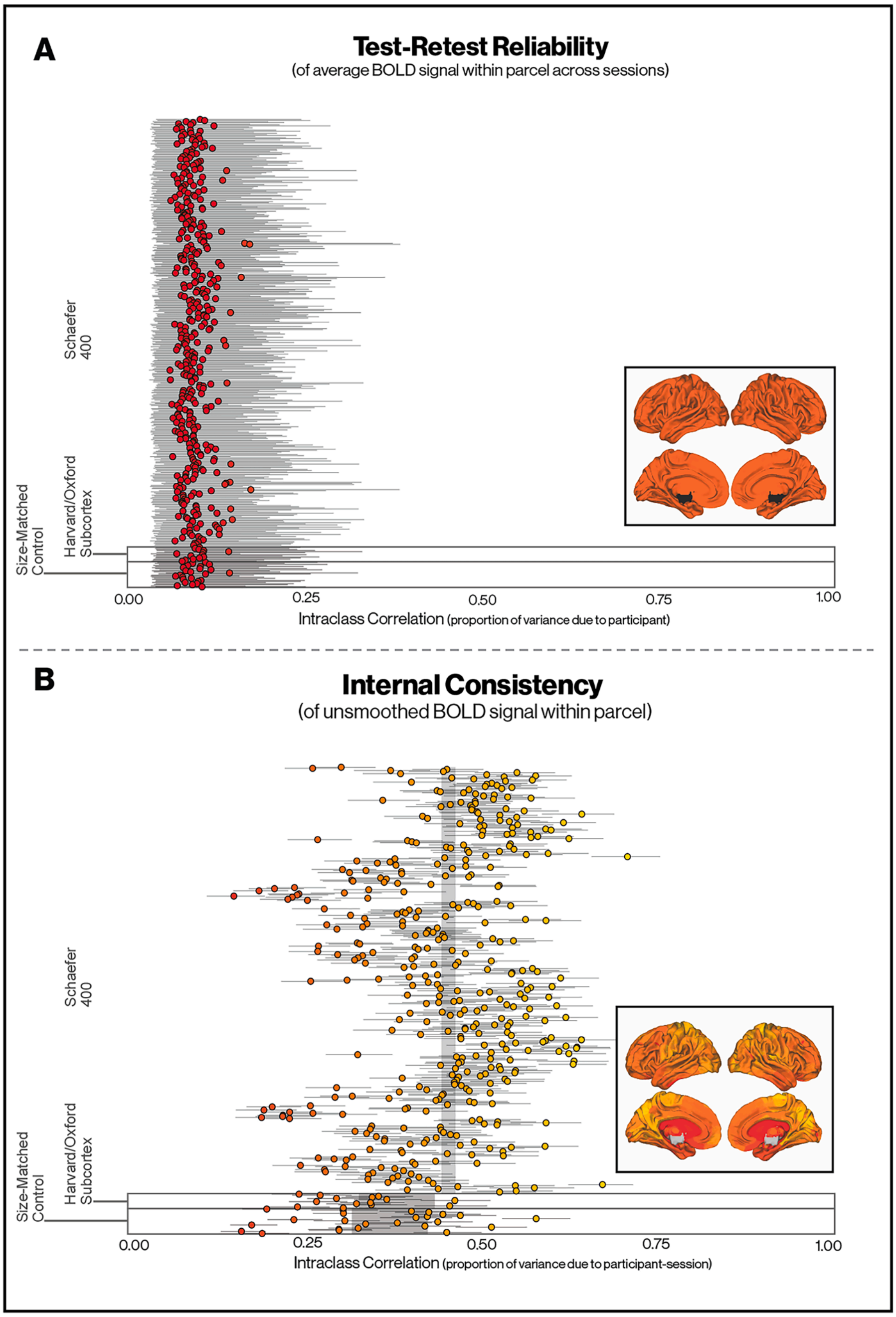

Fig. 3. Test-Retest Reliability and Internal Consistency of fMRI response to aversive stimuli. A: Proportion variance due to participant means (test-retest reliability) across 400 cortical parcels,14 subcortical ROIs, and 24 size-matched control regions. Point estimates are filled with a color corresponding to the ICC, which maps onto the parcellated surface (bottom right corner) with whiskers showing 95% credible intervals. B: Proportion variance due to participant-session means (internal consistency), annotated as in (A). Shaded region indicates 95% confidence interval of expected value from the best fitting model predicting internal consistency for each anatomy type (see methods).

3.4. Changes in neural response to aversive cues over time

Before examining correlates of within-individual variation in neural response to aversive cues (fearful > neutral faces), we first estimated unconditional growth models across the ten monthly scans. Linear decreases in neural response in this contrast were observed in the ventral visual stream, including fusiform and lateral occipital cortex; superior temporal sulcus; dorsomedial prefrontal cortex (PFC); and right middle frontal gyrus (MFG) and inferior frontal gyrus (IFG) (Fig. 5; Tables S2-S3). This pattern of habituation across monthly sessions is broadly consistent with evidence in the literature for within-session habituation to emotional faces in medial and lateral temporal cortex (Fischer et al., 2003). There were no regions where linear increases in activation over time occurred.

3.5. Predicting within-individual variation in task-related neural activation

Having demonstrated some internal consistency of neural responses to aversive cues, we next attempted to explain the fluctuations in these responses within-individuals over time. Specifically, we examined whether within-individual variation in mood, sleep, and stressful life events across time were associated with variability in neural response to aversive stimuli (i.e., fearful > neutral faces) in voxel-wise multilevel

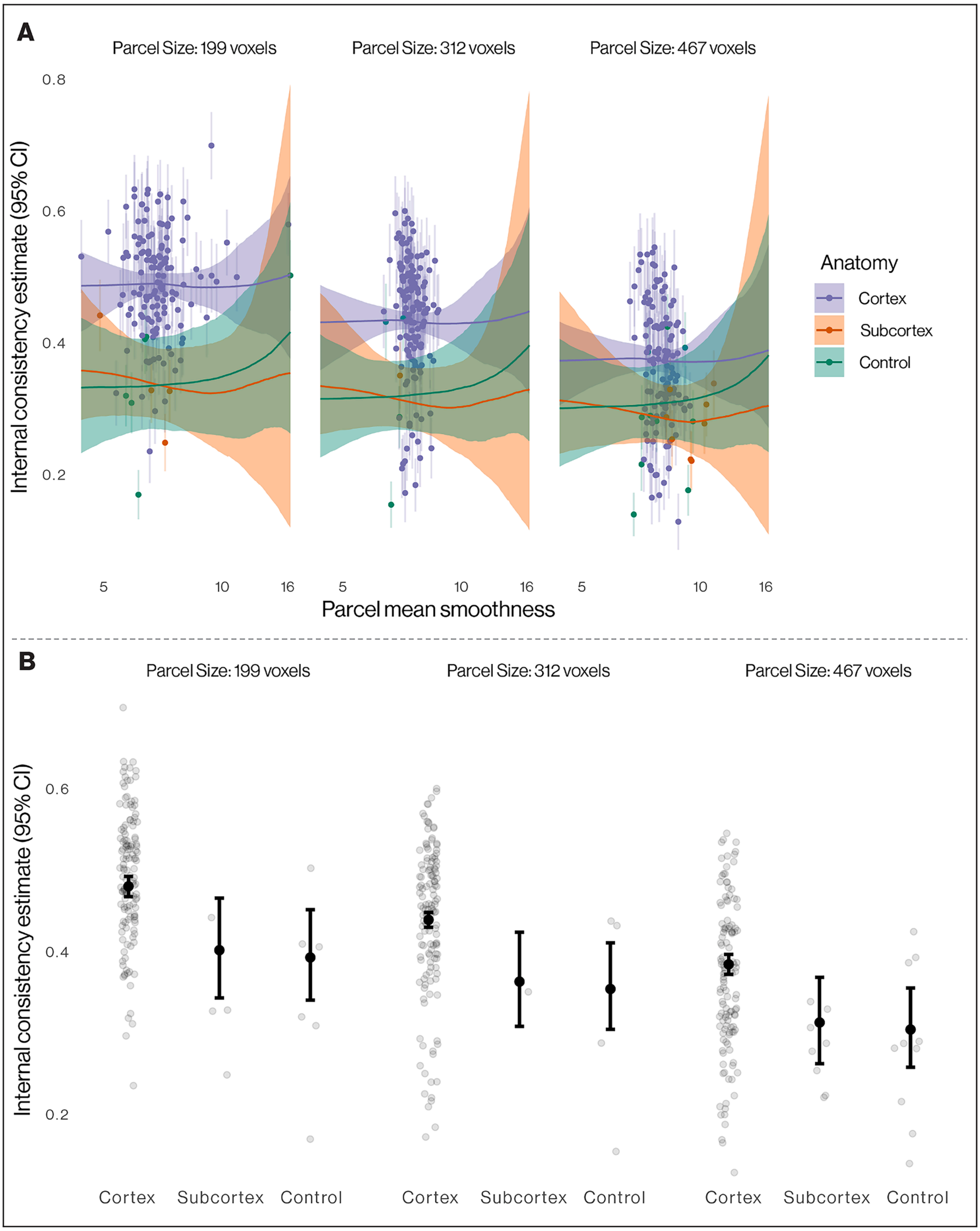

Fig. 4. Effect of smoothness, size, and anatomy on internal consistency. Each panel shows model-expectations when the parcel size is as indicated in the panel titles, corresponding to the 1/6, 3/6, and 5/6 quantiles. Overlaid data come from parcels with sizes in a range centered on those quantiles (i.e., within the first, second, and third third of the data). A: Association between smoothness and internal consistency in the fully unconstrained model. The x-axis has been rescaled to foreground regions with the highest density of data. Whiskers on data points show the 95% credible interval for internal consistency estimates. B: Model expectations from the best fitting model highlighting expected differences in internal consistency between anatomy types. Cortex: Schaefer 400 parcels; subcortex: Harvard-Oxford subcortical regions; control: size-matched control regions.

models that partitioned variance in stress into within-individual and between-individual components (see Methods for details). These variables were not highly correlated within-person ($\rho = [-.11, .25]$; see Table S4 for the full between- and within-person correlation table), so we did not attempt to isolate the unique effect of each variable while controlling for the others. Clusters reflect regions of the brain where monthly fluctuation in the predictor (i.e., mood, sleep, or stress) is associated with corresponding within-individual change in neural response to aversive cues (fearful > neutral faces). Across all analyses, parcel-level analyses seemed more sensitive with more widespread associations than in whole-brain analysis with cluster correction.

Mood. The PANAS demonstrated good internal consistency in this

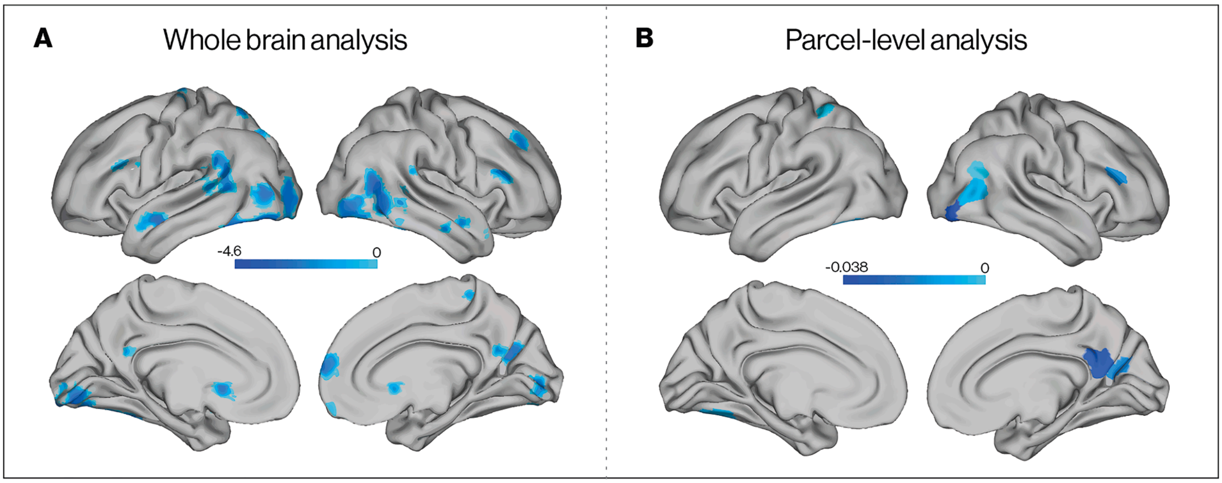

Fig. 5. Effect of time for contrast of Fear > Neutral. A: Statistical map represents t-scores for the effect of study month on the contrast between viewing Fear faces versus Neutral faces; values are plotted for voxels within clusters determined to be significant. B: Parcels with 99.988% credible intervals excluding 0 are displayed using colors representing the effect size for each parcel in terms of standard deviations of the parcel-level outcome for a unit change in the predictor on its original scale.

sample across months; with Chronbach's $\alpha = [.78, .92]$ (IQR = .81 - .90) across months for negative affect. The ICC(1,1) for negative affect was 0.44 and for positive affect was 0.62, indicating that the majority of variance in negative affect and more than one-third of the variance in positive affect was attributable to within-individual variance (Fig. 6).

Monthly fluctuations in negative affect were related to within-individual variation in neural responses to aversive stimuli. On months when participants reported higher negative affect than was typical for them, neural responses to aversive cues (fearful > neutral faces) were lower in a small cluster in IFG in whole-brain analysis (Fig. 7A; Table S5). In parcel-level analysis, neural responses to aversive cues were lower in right IFG, MFG, and temporal-parietal-occipital

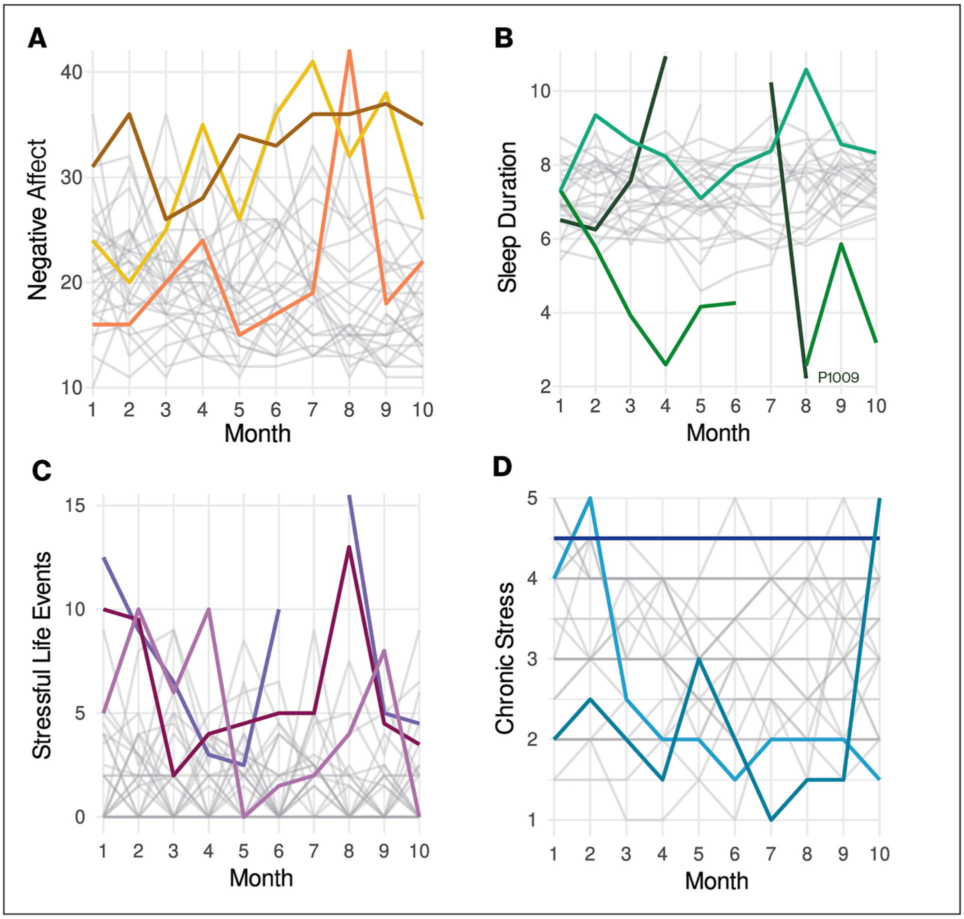

Fig. 6. Variation in predictor variables over time. A value for each measurement, for each participant across all 10 months are shown, with separate lines for each participant. Specific participants are highlighted to illustrate examples with high variability or extreme mean values. Discontinuous lines reflect missing data.

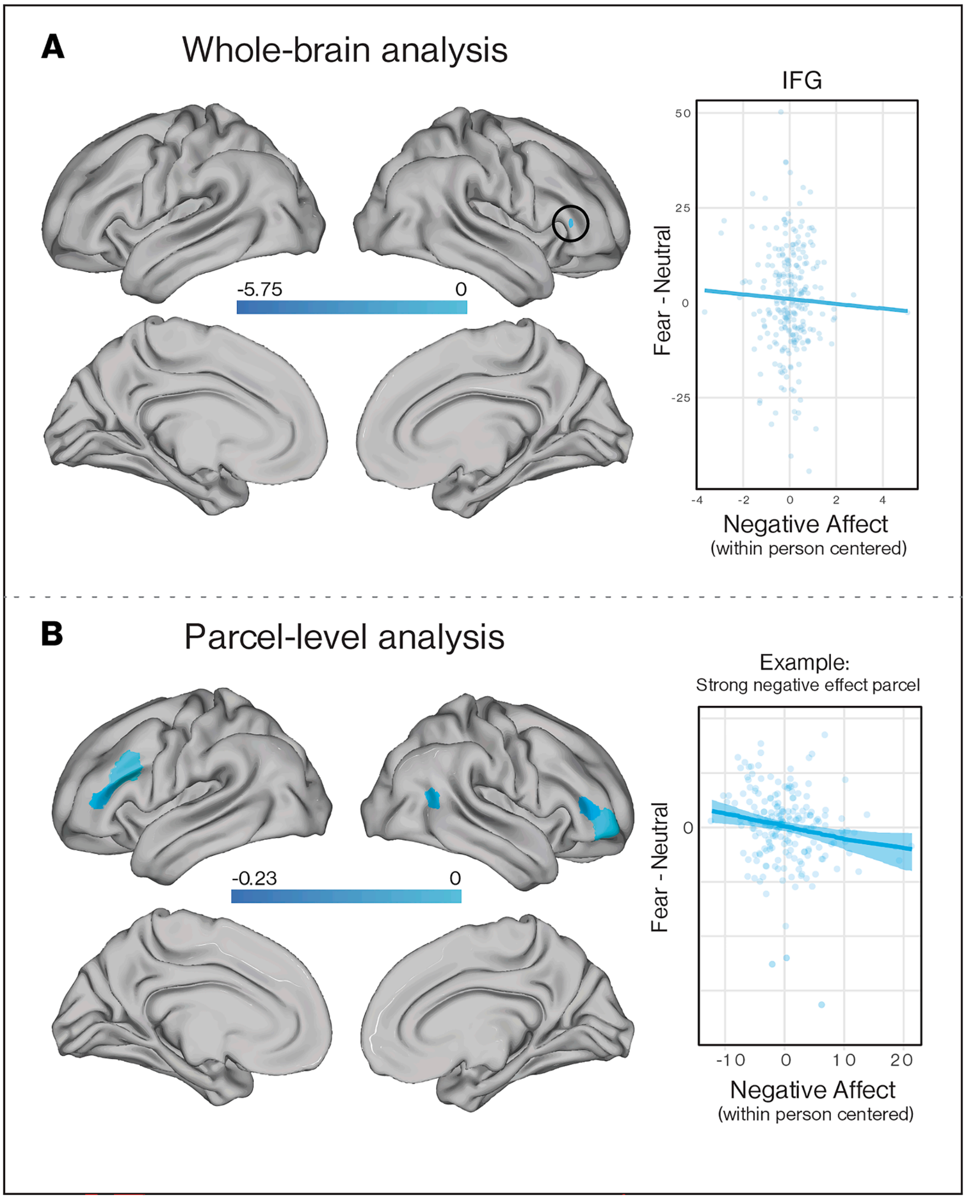

Fig. 7. Associations of monthly within-individual fluctuations in negative affect with within-individual variation in neural response to aversive cues. To conduct within-individual analyses, monthly values of each predictor were first centered around each participant's annual mean; and these annual means, centered at the group mean, were included as a covariate; this approach separates variance at the within-person level from between-person variance. All analyses include month number as a covariate. A: statistical map colors represent t-scores for each effect; values are plotted for voxels within clusters determined to be significant. Scatterplot for the cluster shows mean contrast values on the y axis with the predictor variable on the x axis. The line is the expectation from a model based on mean ROI estimates. B: Parcels with 99.988% credible intervals excluding 0 are displayed using colors representing the effect size for each parcel in terms of standard deviations of the parcel-level outcome for a unit change in the predictor on its original scale. Scatterplots for the example parcels show points for 15 voxels per participant-session with a line for the median of the posterior linear prediction surrounded by the 95% credible interval.

junction (Fig. 7B; Table S6).

Sleep. Sleep duration varied widely over time within-individuals. A majority of the variance in sleep duration occurred within-individuals when examined at the daily level (ICC=0.12) as well as when aggregated across the two-week period prior to each scan (ICC=0.38) (Fig. 6).

Within-individual variation in sleep duration was measured objectively using actigraphy in the two weeks preceding the scan was also related to within-individual variability in neural activation. On months characterized by less sleep than usual, participants exhibited lower activation in three prefrontal clusters spanning right (MFG, frontal pole, and frontal orbital cortex in response to aversive stimuli (fearful > neutral faces) in voxel-wise analysis (Fig. 8A, Table S7). In parcel-level analysis, less sleep than usual was similarly associated with reduced activation in right MFG and frontal pole and higher activation in superior temporal sulcus, precuneus, cuneus, and precentral and postcentral gyrus (Fig. 8B, Table S8).

Stress. The ICC for SLEs was 0.25 and 0.70 for chronic stressors, indicating that the majority of variance in exposure to SLEs and about one-third of the variance in chronic stress occurs within-individuals (Fig. 6). The within-individual correlation between SLEs and chronic stressors was r_{within}= .25, $p < .001$.

Monthly fluctuations in both acute SLEs and chronic stressors—assessed using gold-standard interviews—were similarly related to within-individual variation in neural responses to aversive cues. In voxel-wise analysis, adolescents exhibited heightened neural response to aversive stimuli (fearful > neutral faces) in a cluster spanning

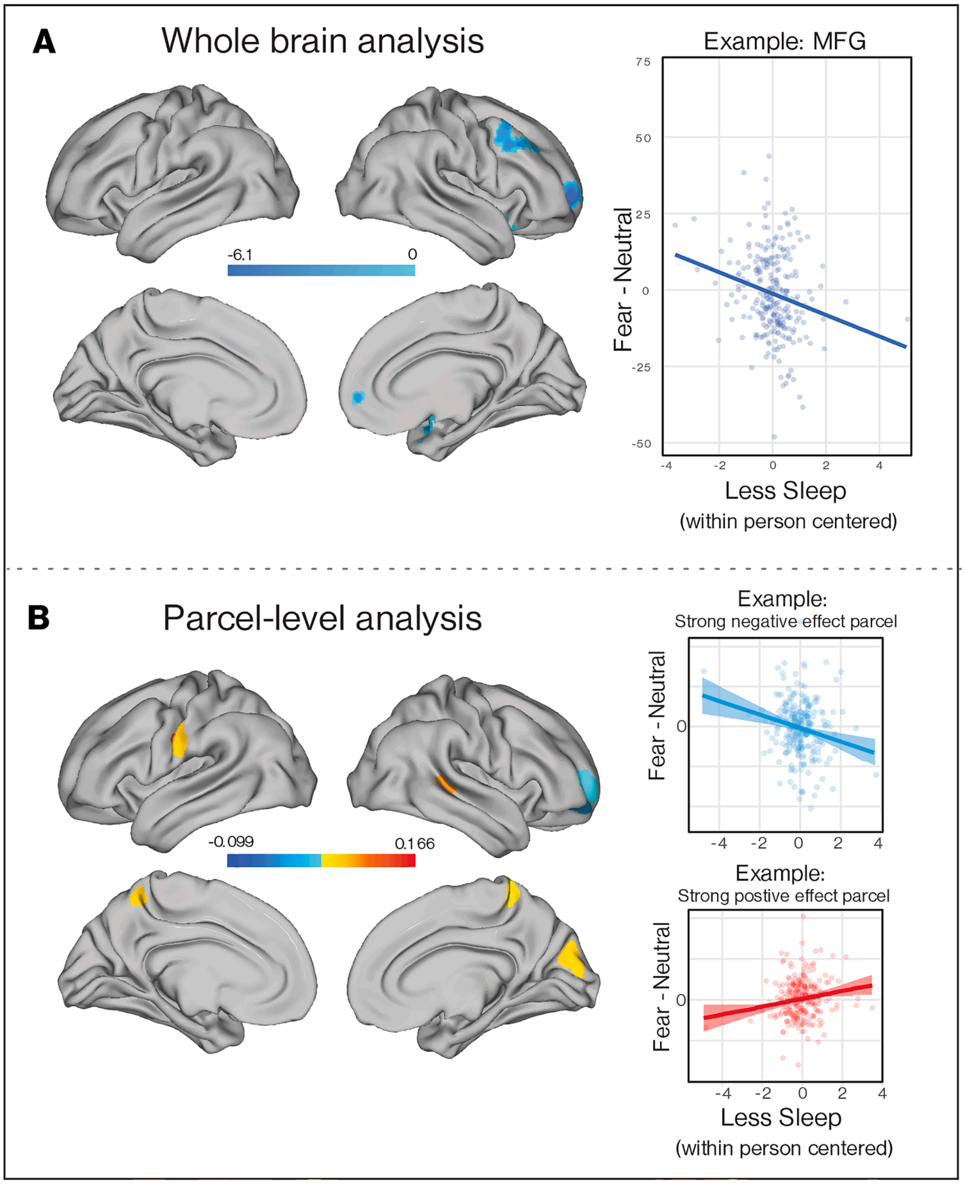

Fig. 8. **Associations of monthly within-individual fluctuations in sleep duration and within-individual variation in neural response to aversive cues.** To conduct within-individual analyses, monthly values of each predictor were first centered around each participant's annual mean and these annual means, centered at the group mean, were included as a covariate; this approach separates variance at the within-person level from between-person variance. Effect of (shorter) sleep duration on Fear > Neutral contrast. All analyses include month number as a covariate. A: statistical map colors represent t-scores for each effect; values are plotted for voxels within clusters determined to be significant. Scatterplot for the example cluster shows mean contrast values on the y axis with the predictor variable on the x axis. The line is the expectation from a model based on mean ROI estimates. B: Parcels with 99.988% credible intervals excluding 0 are displayed using colors representing the effect size for each parcel in terms of standard deviations of the parcel-level outcome for a unit change in the predictor on its original scale. Scatterplots for the example parcels show points for 15 voxels per participant-session with a line for the median of the posterior linear prediction surrounded by the 95% credible interval.

bilateral PCC and precuneus and reduced neural response in a cluster encompassing bilateral dorsal ACC and dorsomedial PFC on months when they experienced more acute SLEs than was typical for them (Fig. 9A, Table S9). These same clusters were observed in parcel-level analysis, as well as reduced activation in bilateral insula, left superior parietal cortex, and left temporoparietal junction and greater activity in bilateral superior frontal gyrus and lateral inferior temporal cortex (Fig. 9B, Table S10).

For chronic stress, we observed a similar pattern of increased neural response to aversive stimuli (fearful > neutral faces) in one cluster in bilateral precuneus on months when adolescents experienced greater chronic stress than usual. We additionally observed reduced within-individual neural response in a cluster spanning right IFG and MFG and in right putamen on months characterized by higher chronic stress than usual (Fig. 9C, Table S11). The precuneus cluster was also observed in parcel-level analysis, along with elevated activation in left cuneus and superior frontal gyrus (Fig. 9D and Table S12).

Amygdala. Left amygdala was credibly associated with the linear effect of time (Fig. S6). The 95% credible intervals included zero for the association between each other predictor and left and right amygdala activity during emotion processing.

4. Discussion

Leveraging a unique sample of adolescents scanned monthly across one year, we investigated the reliability and internal consistency of

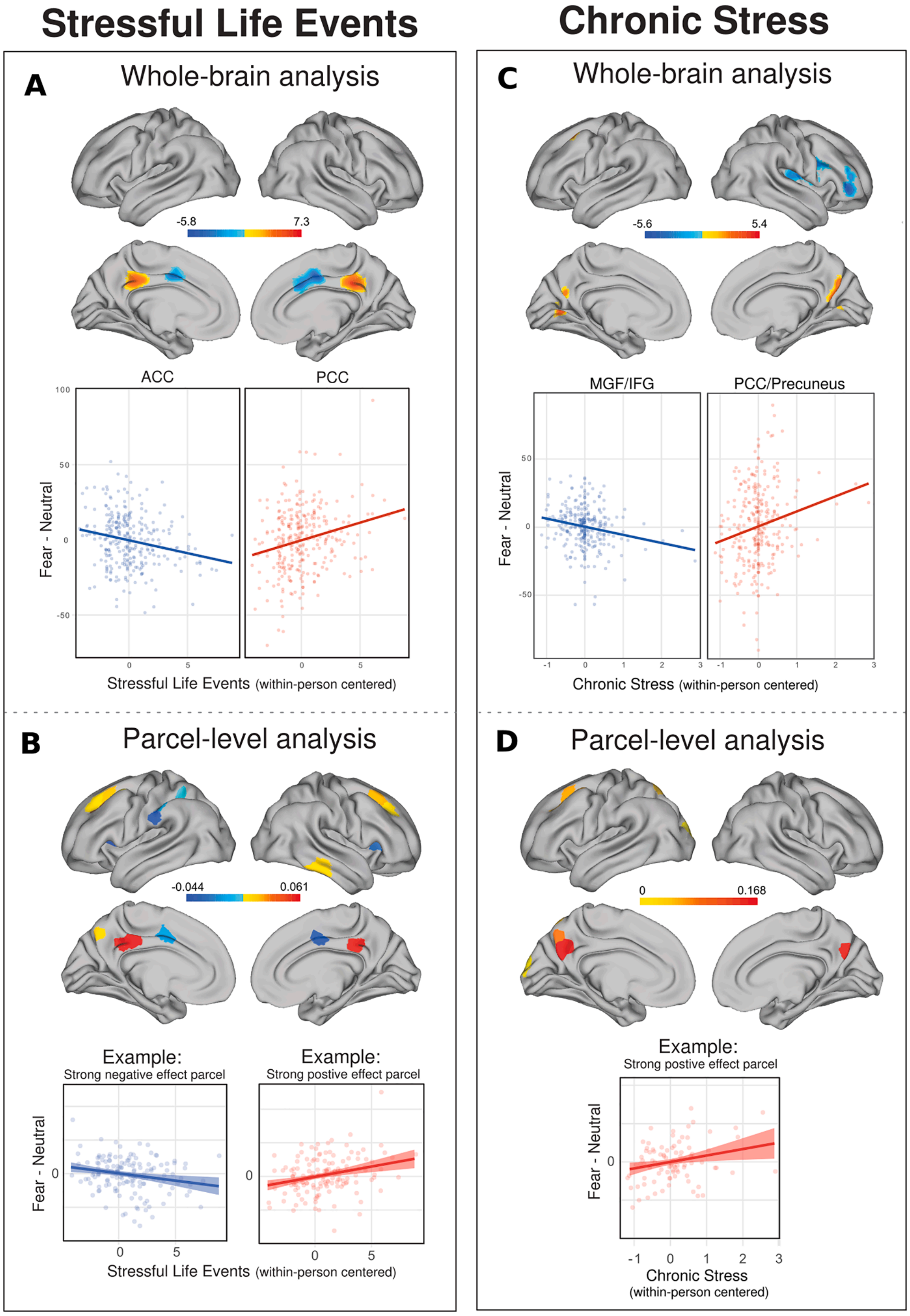

Fig. 9. Associations of monthly within-individual fluctuations in stressors and within-individual variation in neural response to aversive cues. To conduct within-individual analyses, monthly values of each predictor were first centered around each participant's annual mean and these annual means, centered at the group mean, were included as a covariate; this approach separates variance at the within-person level from between-person variance. A, B: Associations of stressful life events with within-individual changes in neural responses in the Fear > Neutral contrast; C, D: Associations of chronic stress with within-individual changes in neural responses in the on Fear > Neutral contrast. All analyses include month number as a covariate. A, C: statistical map colors represent t-scores for each effect; values are plotted for voxels within clusters determined to be significant. Scatterplot for the example cluster shows mean contrast values on the y axis with the predictor variable on the x axis. The line is the expectation from a model based on mean ROI estimates. B, D: Parcels with 99.988% credible intervals excluding 0 are displayed using colors representing the effect size for each parcel in terms of standard deviations of the parcel-level outcome for a unit change in the predictor on its original scale. Scatterplots for the example parcels show points for 15 voxels per participant-session with a line for the median of the posterior linear prediction surrounded by the 95% credible interval.

neural responses during emotion processing over time and within individuals. Although the test-retest reliability of neural responses across time was quite low, internal consistency of BOLD signal across voxels was substantially higher, though this is tempered somewhat by the substantial estimates found in size-matched control regions where no construct-valid variance is expected. Although the ICCs for internal consistency were somewhat lower than standards set for other types of assessments (e.g., interviews, surveys), the estimates obtained here are notable given that they reflect proxy measurement of a complex biological system. Virtually none of the variation in BOLD signal over time reflected stable between-individual differences, whereas nearly half of the variance in cortical parcels reflected internally consistent variance within participants, within sessions, compared to one-third in size-matched control regions. These patterns suggest that instead of reflecting trait-like differences across people, neural responses to aversive cues demonstrate high within-individual variation over time, as well as substantial measurement error when considering the parcel as the unit of measurement. However, while internal consistency was higher in cortical parcels than size-matched control regions, subcortical regions were similar to control regions, suggesting that task response may not be driving signal coherence in subcortical regions. We further demonstrate that this within-individual variability is associated with multiple factors that fluctuate meaningfully over time within individuals. Specifically, within-individual variation in mood, sleep, and exposure to SLEs was associated with dynamic monthly changes in brain activity during emotion processing. These rapid, spatially consistent changes in brain activity suggest that within-individual variation in psychological and physiological states as well as environmental experiences dynamically influence brain activity during adolescence. More broadly, these findings add to growing evidence that precision neuroscience methods have the power to reveal properties of brain function that are not apparent in cross-sectional, between-individual approaches.

The test-retest reliability of BOLD signal in response to aversive cues across the ten neuroimaging sessions was low, indicating poor test-retest reliability of neural responses during emotion processing. This pattern contrasts with two prior reports documenting high reliability in regions of interest for activation during similar emotion processing tasks across two sessions (Gee et al., 2015; Haller et al., 2018). These findings are consistent, however, with a recent meta-analysis and analysis of large cohorts documenting low test-retest reliability of neural response to a range of fMRI tasks, with affective processing tasks exhibiting the lowest reliability (Elliott et al., 2020), and additionally extend the finding of low test-retest reliability to a much shorter time-scale (1 month). These findings have broad implications for affective neuroscience, as these types of emotion processing tasks have been frequently used to study differences in brain function in relation to psychopathology, personality, environmental experiences, and genetics (Canli et al., 2001; Etkin and Wager, 2007; Gray et al., 2005; Groenewold et al., 2013; Hariri et al., 2002; McCrory et al., 2011; Monk et al., 2008; Thomas et al., 2001; Tottenham et al., 2011). Together with the Elliot et al meta-analysis (Elliott et al., 2020), our results suggest that neural responses to affectively-salient cues are poor candidates for research on brain-based biomarkers of stable between-person, i.e., interindividual, variation.

The degree to which this low test-retest reliability extends to neural responses to other types of tasks and metrics of brain activity is an important question for future research. Higher test-retest reliability estimates were observed in the recent meta-analysis for tasks tapping motor and sensory function as well as working memory (Elliott et al., 2020), suggesting that test-retest reliability varies meaningfully as a function of task design and processing domain. Indeed, multivariate patterns of neural response to task demands identified using machine learning may be more stable over time than responses in individual brain areas (Kragel et al., 2021). Functional connectivity of cortical networks assessed using resting-state fMRI may be better suited to studying between-person differences. Although the topography of resting-state cortical networks is highly variable across people (Braga and Buckner, 2017; Gordon et al., 2017), network organization and functional connectivity is highly stable within-individuals over time, with most variance due to stable differences across people (Gratton et al., 2018; Laumann et al., 2017). On the other hand, recent work suggests that the magnitude of brain-behavior associations using resting-state functional connectivity metrics are small and require thousands of individuals to be identified reliably (Marek et al., 2022).

Our findings suggest that poor test-retest reliability of task-evoked BOLD signal likely reflects both high measurement error and high within-individual variability. About half of the variance in cortical parcels and two-thirds in subcortical regions reflects measurement error. The internal consistency of neural response in subcortical regions was only marginally higher than in size-matched control regions in white matter, and ventricles. Virtually all of the remaining variance reflects reliable within-individual variation in the BOLD over time for neural responses during emotional processing. This raises questions about a number of apparently replicable findings that have emerged from between-individual studies of brain activation in these types of emotion processing tasks. For example, elevated neural response in the amygdala and salience network to aversive cues has frequently been observed among people with depression and anxiety disorders (Monk et al., 2008; Swartz et al., 2015b; Thomas et al., 2001) and those who have experienced childhood trauma (McLaughlin et al., 2019). These findings raise the intriguing possibility that rather than reflecting trait-like variability as a function of psychopathology or early-life experiences, these individual differences in neural responses to affective cues may instead reflect state-like factors that vary consistently as a function of psychopathology or exposure to trauma, such as arousal, affect, concentration, sleep, physical activity, or exposure to recent stressors. Future research utilizing dense sampling from the same individuals is needed to explore this possibility.

One solution to dealing with the high measurement error and small effect sizes for brain-behavior associations is to use enormous sample sizes, but another is to utilize the type of densely-sampled longitudinal data of the type we present here (Marek et al., 2022). Within-individual analysis combined with data reduction using parcellations seems to provide good sensitivity in this case. While there is very little reliable between-person variability in neural responses during emotion processing, there is internally consistent systematic variability in BOLD signal fluctuations over time within-individuals. In addition to the methods used here to increase the signal of constructs of interest, this is a potent reminder that there are many other ways to improve signal detection aside from increasing sample size. Still another approach that holds promise in that regard is using individual-level parcellations of network organization given notable individual differences in network topography (Braga and Buckner, 2017; Gordon et al., 2017). Indeed, recent work suggests that brain-behavior associations are larger when individual-specific parcellations are used (Kong et al., 2021, 2019).

It is important to note that there are many possible sources of state variance in the BOLD signal aside from task-related neural activity. Even though covariates can reduce many of these sources, it is not possible to completely eliminate them. For example, BOLD signal may be influenced by heartbeat, respiration, hydration, and motion (Liu, 2016) that are idiosyncratic to a particular person during a particular session, but spatially coherent across voxels in a parcel. Other properties of the MR signal may also contribute to internal consistency of BOLD signal within parcels, for example, aspects of the scanner environment like temperature, humidity, or how many scans were done previously. In the present study, we computed internal consistency for the contrast of fearful relative to neutral faces. The above state-like sources of noise, if they are fairly stable across the duration of the task, are likely somewhat ameliorated when we subtract the signal during neutral blocks from signal during fear blocks. In addition, if spatial coherence in the BOLD signal were primarily responsible for the internal consistency we see within-individuals, we would expect to observe a positive association between smoothness and internal consistency. In contrast, smoothness was unrelated to these estimates. However, the above does not address the broader question of validity. Indeed, the internal consistency of size-matched control regions was non-trivial, indicating that there are sources of spatially coherent but invalid signal. As such, we view the estimates of internal consistency as upper bounds of the true reliability of task-evoked BOLD signal for emotion processing tasks, and certainly as a generous upper-bound on validity.

Importantly, we demonstrate that monthly fluctuations in mood, sleep, and exposure to stress predict variation in neural responses to aversive cues over time within individuals. Notable convergence was observed across these predictors. Adolescents exhibited decreased activation in dorsal and ventral lateral PFC in response to aversive stimuli on months when they reported lower mood, had less sleep, and experienced higher exposure to SLEs than was typical for them; we also observed decreased activation in dorsal ACC on months characterized by less sleep and higher than usual stress. These results demonstrate consistent within-individual reductions in neural response to aversive cues in regions involved in monitoring control-relevant information (e.g., conflict) and implementing control processes (Shenhav et al., 2013). Recent evidence demonstrates shared neural representation of aversive stimuli and cognitive conflict in dorsal ACC, which suggests that aversive stimuli signal control demands similarly to conflict (Vermeylen et al., 2020). Given the absence of a meaningful behavioral response in this task, it is unclear what these neural patterns might reflect (Poldrack, 2011), although it is worth noting that these PFC regions are also recruited in many forms of emotional processing and emotion regulation (Buhle et al., 2014; Etkin et al., 2011). Finally, we observed within-individual increases in activation in PCC and precuneus—key nodes in the default network—in response to aversive stimuli on months characterized by higher levels of SLEs and chronic stress and, to a lesser extent, higher negative mood. PCC activation occurs in response to positive and negative affective stimuli (Maddock et al., 2003), when reflecting on emotional states in oneself and others (Ochsner et al., 2004), and in self-referential processing (Northoff et al., 2006), including auto-biographical memory (Sugiura et al., 2005). It is possible that aversive stimuli are more likely to trigger self-focused thinking, such as rumination, during periods characterized by high levels of stress. Indeed, exposure to SLEs is associated with increased engagement in rumination in longitudinal studies (Michl et al., 2013; Moberly and Watkins, 2008). However, interpretation of these neural patterns is speculative.

To our knowledge, these types of within-individual associations with neural activity have not previously been documented. These patterns add to growing evidence that precision neuroscience, which focuses on repeated sampling of the same individuals, may stimulate progress in characterizing individual variation in brain function. While task-related BOLD signal—at least in response to affective cues—is likely a poor candidate for studying stable individual variation, as some have recently argued (Elliott et al., 2020), it may still be suited to studying changes in brain function within-individual variation over time. Indeed, we show that fluctuations in neural responses over time are associated with relevant processes known to vary over time within-individuals—including mood, sleep, and exposure to SLEs. These findings parallel recent work in psychology demonstrating high within-individual variation in a range of constructs that have historically been studied in between-person designs—including affect, psychopathology symptoms, and physiological markers—and poor correspondence between associations observed in between-participant designs from those derived from within-participant longitudinal studies (Fisher et al., 2018). Our findings suggest that intensive longitudinal designs that probe neural function in the same individuals over time are needed to characterize this within-individual variability, its correlates, and important measurement characteristics like reliability with more focus on the temporal resolution of the underlying constructs.

We focus here on neural responses to affectively-salient stimuli. Affective constructs are known to have high within-individual variability (Fisher et al., 2018). However, even constructs that are more trait-like—such as working memory—show meaningful variation within-individuals over time that are strongly linked to fluctuations, for example, in stress (Sliwinski et al., 2006). Although we cannot extrapolate from the constructs measured here, it is plausible that neural circuits that support a range of cognitive and affective constructs frequently studied in cognitive neuroscience might exhibit meaningful within-individual variability.

The attention to within-individual variation highlights a broader conceptual point about how cognitive neuroscientists design and analyze studies. Observing associations between constructs over time can be an important tool for unraveling causal effects (Collins, 2006; Raudenbush, 2001; Rohrer and Murayama, 2023). Moreover, the degree of variation over time can itself be an important predictor or outcome, as prior research has shown, for example, in sleep (Vidal Bustamante et al., 2020), and emotion perception and experience (Nook et al., 2021; O'Toole et al., 2020). Even in cross-sectional studies, investigators often have repeated measurements, and appropriately interrogating the inherent within-individual variation has been shown to be beneficial for obtaining reliable measurement of selective attention in the Stroop task (Haines et al., 2020). Designing studies to reap the numerous benefits of within-individual variability will fortify the empirical value of our science.

4.1. Limitations

There are several aspects of the current study design and set of analyses that should be considered in interpreting these findings. We were not able to address several potential causes of variance in reliability and internal consistency here either because of data limitations or scope. Most importantly, as stated in the introduction, is the fact that we examine reliability and internal consistency, not validity. While reliability is necessary for validity, it is not sufficient.

The particulars of a given parcellation scheme or set of ROIs may impact our estimates. Different parcellations have been shown to affect measurements of functional activity and brain-behavior associations, and the results here would be expected to be influenced by parcellation choice as well (Bryce et al., 2021; Flournoy et al., 2020). We focus on subcortical structures and a widely-used cortical parcellation that divides the cortex into 400 regions based on patterns of task-related activation and functional connectivity at rest (Schaefer et al., 2018). Our estimates, especially for internal consistency, would likely be higher for smaller parcels, and lower as parcel size increases, although registration error might contravene this expectation. Indeed, parcellations based on within-person functional connectivity would likely yield the highest signal reliability (Kong et al., 2019; Xue et al., 2021). The goal here is not to tie our findings to a particular parcellation scheme, but rather to examine internal consistency of BOLD signal to explore to what extent measurement error causes low test-retest reliability across the month-long intervals examined.

We were not able to evaluate test-retest reliability at other timescales, which would be an important next step in corroborating the conclusions drawn from the internal consistency results. Such analyses should first consider the timescale over which test-retest reliability in the particular sensory, cognitive, or affective process in question would be expected. At a minimum, higher test-retest reliability would be expected at extremely short time intervals, with the caveat that well-known habituation effects would need to be accounted for as well. With regard to the behavioral variables examined here, monthly variation has been used in prior studies of within-individual variation in stressful life events and sleep (Nook et al., 2021; Rodman et al., 2021; Vidal Bustamante et al., 2020). Assessing stressful life events at shorter intervals, particularly using the objective interview-based approach utilized here, is not feasible. However, timescales ranging from days to weeks might be more appropriate for capturing meaningful within-individual variation in mood. Additional research is needed to determine the appropriate timescale for assessing within-individual variation in neural function across different modalities.

This data collection was not designed to test the variability of test-retest reliability or internal consistency across multiple fMRI tasks or methods of analysis. Even within the domain of affect processing variation in test-retest reliability would be expected in different kinds of emotion processes, but we were not able to test these ideas here. This analysis is focused on more traditional analysis methods (contrast-based, GLM task fMRI analysis), that remain dominant in the literature, and thus the results here do not generalize to more modern, multivariate methods (Marek et al., 2020, 2019). In general, multivariate methods have been shown to have higher reliability (Kragel et al., 2021), and larger effect sizes (Marek et al., 2020), possibly because they better take advantage of widely distributed neural representations. We would expect that to hold for these data as well, but produce the same dissociation between test-retest reliability and internal consistency. Investigating the reliability of multivariate methods in multiple ways is an important next step in this line of research.

The degree of reliability and internal consistency we document here should be taken as an approximate measure with some caveats. Reliability is not a property of just of a test, but also of the people taking the test; in this case, we report on a particular fMRI task in particular parcels and ROIs in adolescent females. Researchers sampling from other populations, or using different parcels, or a different task, should not derive but the most vague expectations from these results. Moreover, these estimates were generated using two common approaches to estimating reliability, but there is considerably more work to be done in developing measurement models for neural data.

Finally, while we document some internal consistency, we are only able to provide cursory evidence as to the validity of this signal. Supporting the validity of this fMRI task as a measure of differences in emotional processing, we saw associations with mood, sleep, and multiple sources of stress that were expected based on between-person studies (Arnone et al., 2012; Goldstein-Piekarski et al., 2015; Larson and Ham, 1993; Mroczek and Almeida, 2004; Sliwinski et al., 2009; Swartz et al., 2015b, 2015a; Wang et al., 2006). However, we do not have more proximal measures of emotional processing that would allow us to test convergent validity, nor do we have measures that would help us rule out other sources of variance that are distinct from emotional processing, yet would also be expected to correlate with neural responses to this task, such as arousal or attention (Zelkowitz and Cole, 2016). This is, perhaps, the more pressing question for cognitive neuroscience at this point. Validity work is often less of a focus in behavioral tasks than surveys (Clark and Watson, 2019), but is no less important (see Schiavone et al., 2023 for a tool to evaluate multiple validities). We often rely on face validity, but this ignores the possibility that variations in task performance and neural responses may be caused by related but distinct processes. This is an important consideration for a precision-neuroscience approach that seeks to discover biomarkers, and to lead us toward mechanistic explanations of these processes.

4.2. Conclusion

Leveraging an intensive longitudinal study with 10 monthly scans per participant, we observe low test-retest reliability of neural responses to aversive cues over time. Using a common approach to assessing internal consistency, we demonstrate that measurement of BOLD signal related to emotion processing is moderately consistent within cortical regions, suggesting that this temporal instability across months may in part reflect high within-individual variation in neural responses to affectively-salient cues in addition to measurement error. Internally consistent within-individual variation accounted for roughly half of the variance in BOLD signal over time in cortical parcels and a third in both subcortical regions and size matched control regions, whereas between-person differences explained virtually none. Within-individual variation in sleep, mood, and stress all contributed to this within-individual variability in neural responses. These findings highlight the importance of evaluating the test-retest reliability of neural responses to other types of fMRI tasks to ensure that fMRI studies are designed in a way that accurately reflects the underlying temporal properties of the construct being measured. Doing so could bring needed nuance to discussions of the validity and reliability of task fMRI data for studying individual variation in brain function.

CRediT authorship contribution statement

John C. Flournoy: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – review & editing. **Nessa V. Bryce:** Conceptualization, Visualization, Writing – review & editing. **Meg J. Dennison:** Investigation, Writing – review & editing. **Alexandra M. Rodman:** Investigation, Writing – review & editing. **Elizabeth A. McNeilly:** Investigation, Writing – review & editing. **Lucy A. Lurie:** Investigation, Writing – review & editing. **Debbie Bitran:** Investigation, Writing – review & editing. **Azure Reid-Russell:** Investigation, Writing – review & editing. **Constanza M. Vidal Bustamante:** Investigation, Formal analysis, Methodology, Writing – review & editing. **Tara Madhyastha:** Conceptualization, Data curation, Software, Writing – review & editing. **Katie A. McLaughlin:** Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

None.

Data availability

Code and data repository: https://osf.io/zy92w/.

Acknowledgements

We are deeply grateful for the 30 adolescent girls who gave their time month by month over the course of a year to make this research possible. Thanks to Debbie Bitran, Lucy Lurie, Elizabeth McNeilly, Zoe Miles, and Azure Reid-Russell for participant recruitment and testing. Thanks to Randy Buckner, Gang Chen, Eveline Crone, Max Elliot, Ted Satterthwaite, Leah Somerville, Neuroimage Editors, and Reviewer 2 for input on the manuscript.

Funding

The study was funded by grants from the One Mind Institute, AIM for Mental Health, and the National Institute of Mental Health (R37-MH119194) to KM, K99/R00-MH126163 to AR.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2023.120503.

References

Arnone, D., McKie, S., Elliott, R., Thomas, E.J., Downey, D., Juhasz, G., Anderson, I.M., 2012. Increased amygdala responses to sad but not fearful faces in major depression: relation to mood state and pharmacological treatment. Am. J. Psychiatry 169, 841–850.

Barch, D.M., Burgess, G.C., Harms, M.P., Petersen, S.E., Schlaggar, B.L., Corbetta, M., Glasser, M.F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J.M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A.Z., Van Essen, D.C., 2013. Function in the human connectome: Task-fMRI and individual differences in behavior. NeuroImage 80, 169–189. https://doi.org/10.1016/j.neuroimage.2013.05.033.

Bastiaansen, J.A., Bennik, E.C., Marsman, J.B.C., Ormel, J., Aleman, A., Oldehinkel, A.J., 2018. Prefrontal cortex activation during a cognitive reappraisal task is associated with real-life negative affect reactivity. PLOS One 13, e0202888. https://doi.org/10.1371/journal.pone.0202888.

Behzadi, Y., Restom, K., Liau, J., Liu, T.T., 2007. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. Neuroimage 37, 90–101.

Bonito, J.A., Ruppel, E.K., Keyton, J., 2012. Reliability estimates for multilevel designs in group research. Small Group Res. 43, 443–467. https://doi.org/10.1177/1046496412437614.

Braams, B.R., van Duijvenvoorde, A.C.K., 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, 7226–7238. https://doi.org/10.1523/JNEUROSCI.4764-14.2015.

Braga, R.M., Buckner, R.L., 2017. Parallel interdigitated distributed networks within the individual estimated by intrinsic functional connectivity. Neuron 95, 457–471.

Breiter, H.C., Etcoff, N.L., Whalen, P.J., Kennedy, W.A., Rauch, S.L., Buckner, R.L., Strauss, M.M., Hyman, S.E., Rosen, B.R., 1996. Response and habituation of the human amygdala during visual processing of facial expression. Neuron 17, 875–887.

Bryce, N.V., Flournoy, J.C., Guassi Moreira, J.F., Rosen, M.L., Sambook, K.A., Mair, P., McLaughlin, K.A., 2021. Brain parcellation selection: an overlooked decision point with meaningful effects on individual differences in resting-state functional connectivity. NeuroImage 243, 118487. https://doi.org/10.1016/j.neuroimage.2021.118487.

Buhle, J.T., Silvers, J.A., Wager, T.D., Lopez, R., Onyemekwu, C., Kober, H., Weber, J., Ochsner, K.N., 2014. Cognitive reappraisal of emotion: a meta-analysis of human neuroimaging studies. Cereb. Cortex 24, 2981–2990. https://doi.org/10.1093/cercor/bht154.

Bürkner, P.-C., 2018. Advanced Bayesian multilevel modeling with the R package brms. R J 10, 395–411.

Bürkner, P.-C., 2017. brms: an R package for bayesian multilevel models using stan. J. Stat. Softw. 80, 1–28. https://doi.org/10.18637/jss.v080.i01.

Canli, T., Zhao, Z., Desmond, J.E., Kang, E., Gross, J., Gabrieli, J.D., 2001. An fMRI study of personality influences on brain reactivity to emotional stimuli. Behav. Neurosci. 115, 33–42.

Chen, G., Pine, D.S., Brotman, M.A., Smith, A.R., Cox, R.W., Haller, S.P., 2021. Beyond the intraclass correlation: a hierarchical modeling approach to test-retest assessment. bioRxiv 2021.01.04.425305. https://doi.org/10.1101/2021.01.04.425305.

Chen, G., Taylor, P.A., Haller, S.P., Kircanski, K., Stoddard, J., Pine, D.S., Leibenluft, E., Brotman, M.A., Cox, R.W., 2018. Intraclass correlation: improved modeling approaches and applications for neuroimaging. Hum. Brain Mapp. 39, 1187–1206. https://doi.org/10.1002/hbm.23909.

Clark, L.A., Watson, D., 2019. Constructing validity: new developments in creating objective measuring instruments. Psychol. Assess. 31, 1412–1427. https://doi.org/10.1037/pas0000626.

Collins, L.M., 2006. Analysis of longitudinal data: the integration of theoretical model, temporal design, and statistical model. Annu. Rev. Psychol. 57, 505–528. https://doi.org/10.1146/annurev.psych.57.102904.190146.

Cosme, D., Flournoy, J.C., Livingston, J.L., Lieberman, M.D., Dapretto, M., Pfeifer, J.H., 2022. Testing the adolescent social reorientation model during self and other evaluation using hierarchical growth curve modeling with parcellated fMRI data. Dev. Cogn. Neurosci. 54, 101089 https://doi.org/10.1016/j.dcn.2022.101089.

Cox, R.W., 2019. Equitable thresholding and clustering: a novel method for functional magnetic resonance imaging clustering in AFNI. Brain Connect. 9, 529–538. https://doi.org/10.1089/brain.2019.0666.

Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162–173. https://doi.org/10.1006/cbmr.1996.0014.

Cox, R.W., Hyde, J.S., 1997. Software tools for analysis and visualization of fMRI data. NMR Biomed 10, 171–178.doi:10.1002/(SICI)1099-1492(199706/08)10:4/5<171::AID-NBM453>3.0.CO;2-L.

Daley, S.E., Hammen, C., Burge, D., Davila, J., Paley, B., Lindberg, N., Herzberg, D.S., 1997. Predictors of the generation of episodic stress: a longitudinal study of late adolescent women. J. Abnorm. Psychol. 106, 251–259.

de Zambotti, M., Baker, F.C., Willoughby, A.R., Godino, J.G., Wing, D., Patrick, K., Colrain, I.M., 2016. Measures of sleep and cardiac functioning during sleep using a multisensory commercially-available wristband in adolescents. Physiol. Behav. 158, 143–149.

Dohrenwend, B.P., 2006. Inventorying stressful life events as risk factors for psychopathology: Toward resolution of the problem of intracategory variability. Psychol. Bull. 132, 477–495.

Dohrenwend, B.P., Shrout, P.E., 1985. Hassles" in the conceptualization and measurement of life stress variables. Am. Psychol. 40, 780–785.

Elliott, M.L., Knodt, A.R., Ireland, D., Morris, M.L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T.E., Caspi, A., Hariri, A.R., 2020. What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. Psychol. Sci. https://doi.org/10.1177/0956797620916786, 095679762091678.

Enders, C.K., Tofighi, D., 2007. Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. Psychol. Methods 12, 121–138. https://doi.org/10.1037/1082-989X.12.2.121.

Etkin, A., Egner, T., Kalisch, R., 2011. Emotional processing in anterior cingulate and medial prefrontal cortex. Trends Cogn. Sci. 15, 85–93. https://doi.org/10.1016/j.tics.2010.11.004.

Etkin, A., Wager, T.D., 2007. Functional neuroimaging of anxiety: a meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia. Am. J. Psychiatry 164, 1476–1488.

Fischer, H., Wright, C.I., Whalen, P.J., McInerney, S.C., Shin, L.M., Rauch, S.L., 2003. Brain habituation during repeated exposure to fearful and neutral faces: a functional MRI study. Brain Res. Bull. 59, 387–392.

Fisher, A.J., Medaglia, J.D., Jeronimus, B.F., 2018. Lack of group-to-individual generalizability is a threat to human subjects research. Proc. Natl. Acad. Sci. 115, E6106–E6115. https://doi.org/10.1073/pnas.1711978115.

Flournoy, J.C., Vijayakumar, N., Cheng, T.W., Cosme, D., Flannery, J.E., Pfeifer, J.H., 2020. Improving practices and inferences in developmental cognitive neuroscience. Dev. Cogn. Neurosci. 45, 100807 https://doi.org/10.1016/j.dcn.2020.100807.

Forbes, E.E., Phillips, M.L., Silk, J.S., Ryan, N.D., Dahl, R.E., 2011. Neural systems of threat processing in adolescents: role of pubertal maturation and relation to measures of negative affect. Dev. Neuropsychol. 36, 429–452. https://doi.org/10.1080/87565641.2010.550178.

Freedman, D., Lane, D., 1983. A nonstochastic interpretation of reported significance levels. J. Bus. Econ. Stat. 1, 292–298. https://doi.org/10.2307/1391660.

Fusar-Poli, P., Placentino, A., Carletti, F., Allen, P., Landi, P., Abbamonte, M., Barale, F., Perez, J., McGuire, P., Politi, P.L., 2009. Laterality effect on emotional faces processing: ALE meta-analysis of evidence. Neurosci. Lett. 452, 262–267. https://doi.org/10.1016/j.neulet.2009.01.065.

Garrison, J., Erdeniz, B., Done, J., 2013. Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. Neurosci. Biobehav. Rev. 37, 1297–1310.

Gee, D.G., McEwen, S.C., Forsyth, J.K., Haut, K.M., Bearden, C.E., Addington, J., Goodyear, B., Candenhead, K.S., Mirzakhanian, H., Cornblatt, B.A., Olvet, D., Mathalon, D.H., McGlashan, T.H., Perkins, D.O., Belger, A., Seidman, L.J., Thermenos, H., Tsuang, M.T., van Erp, T.G.M., Walker, E.F., Hamann, S., Woods, S. W., Constable, T., Cannon, T.D., 2015. Reliability of an fMRI paradigm for emotional processing in a multisite longitudinal study. Hum. Brain Mapp. 36, 2558–2579.

Gelman, A., Carlin, J., 2014. Beyond power calculations assessing type S (Sign) and Type M (Magnitude) errors. Perspect. Psychol. Sci. 9, 641–651. https://doi.org/10.1177/1745691614551642.

Ghosh, S.S., Kakunoori, S., Augustinack, J., Nieto-Castanon, A., Kovelman, I., Gaab, N., Christodoulou, J.A., Triantafyllou, C., Gabrieli, J.D.E., Fischl, B., 2010. Evaluating the validity of volume-based and surface-based brain image registration for developmental cognitive neuroscience studies in children 4-to-11 years of age. NeuroImage 53, 85–93. https://doi.org/10.1016/j.neuroimage.2010.05.075.

Goldstein-Piekarski, A.N., Greer, S.M., Saletin, J.M., Walker, M.P., 2015. Sleep deprivation impairs the human central and peripheral nervous system discrimination of social threat. J. Neurosci. 35, 10135–10145.

Gonzalez-Castillo, J., Chen, G., Nichols, T.E., Bandettini, P.A., 2017. Variance decomposition for single-subject task-based fMRI activity estimates across many sessions. NeuroImage, Clean. fMRI Time Series: Mitigating Noise Adv. Acquisition Correction Strat. 154, 206–218. https://doi.org/10.1016/j.neuroimage.2016.10.024.

Gordon, E.M., Laumann, T.O., Gilmore, A.W., Newbold, D.J., Greene, D.J., Berg, J.J., Ortega, M., Hoyt-Drazen, C., Gratton, C., Sun, H., Hampton, J.M., Coalson, R.S., Nguyen, A.L., McDermott, K.B., Shimony, J.S., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., Nelson, S.M., Dosenbach, N.U.F., 2017. Precision functional mapping of individual human brains. Neuron 95, 791–807. https://doi.org/10.1016/j.neuron.2017.07.011 e7.

Gratton, C., Kraus, B.T., Greene, D.J., Gordon, E.M., Laumann, T.O., Nelson, S.M., Dosenbach, N.U.F., Petersen, S.E., 2020. Defining individual-specific functional neuroanatomy for precision psychiatry. Biol. Psychiatry, Converg. Heterogen. Psychopathol. 88, 28–39. https://doi.org/10.1016/j.biopsych.2019.10.026.

Gratton, C., Laumann, T.O., Nielsen, A.N., Greene, D.J., Gordon, E.M., Gilmore, A.W., Nelson, S.M., Coalson, R.S., Snyder, A.Z., Schlaggar, B.L., Dosenbach, N.U.F., Petersen, S.E., 2018. Functional brain networks are dominated by stable group and individual factors, not cognitive or daily variation. Neuron 98, 439–452. https://doi.org/10.1016/j.neuron.2018.03.035 e5.

Gray, J.R., Burgess, G.C., Schaefer, A., Yarkoni, T., Larsen, R.J., Braver, T.S., 2005. Affective personality differences in neural processing efficiency confirmed using fMRI. Cogn. Affect. Behav. Neurosci. 5, 182–190.

Groenewold, N.A., Opmeer, E.M., de Jonge, P., Aleman, A., Costafreda, S.G., 2013. Emotional valence modulates brain functional abnormalities in depression: evidence from a meta-analysis of fMRI studies. Neurosci. Biobehav. Rev. 37, 152–163.

Haines, N., Kvam, P.D., Irving, L.H., Smith, C., Beauchaine, T.P., Pitt, M.A., Ahn, W.-Y., Turner, B., 2020. Theoretically informed generative models can advance the psychological and brain sciences: lessons from the reliability paradox (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/xr7y3.

Haller, S.P., Kircanski, K., Stoddard, J., White, L.K., Chen, G., Sharif-Askary, B., Zhang, S., Towbin, K.E., Pine, D.S., Leibenluft, E., Brotman, M.A., 2018. Reliability of neural activation and connectivity during implicit face emotion processing in youth. Dev. Cogn. Neurosci. 31, 67–73.

Hammen, C., 1991. The generation of stress in the course of unipolar depression. J. Abnorm. Psychol. 100, 555–561.

Hammen, C., 1988. Self-cognitions, stressful events, and the prediction of depression in children of depressed mothers. J. Abnorm. Child Psychol. 16, 347–360.

Hammen, C., Henry, R., Daley, S.E., 2000. Depression and sensitization to stressors among young women as a function of childhood adversity. J. Consult. Clin. Psychol. 68, 782–787. https://doi.org/10.1037//0022-006X.68.5.782.

Hankin, B.L., Abramson, L.Y., Moffitt, T.E., Silva, P.A., McGee, R., Angell, K.E., 1998. Development of depression from preadolescence to young adulthood: Emerging gender differences in a 10-year longitudinal study. J. Abnorm. Psychol. 107, 128–140.

Hariri, A.R., Mattay, V.S., Tessitore, A., Kolachana, B., Fera, F., Goldman, D., Weinberger, D.R., 2002. Serotonin transporter genetic variation and the response of the human amygdala. Science 297, 400–403.

Huang, C.-M., Lee, S.-H., Hsiao, I.-T., Kuan, W.-C., Wai, Y.-Y., Ko, H.-J., Wan, Y.-L., Hsu, Y.-Y., Liu, H.-L., 2010. Study-specific EPI template improves group analysis in functional MRI of young and older adults. J. Neurosci. Methods 189, 257–266. https://doi.org/10.1016/j.jneumeth.2010.03.021.

Huettel, S.A., Song, A.W., McCarthy, G., 2004. Functional Magnetic Resonance Imaging. Sinauer Associates, Sunderland, MA.

Kong, R., Li, J., Orban, C., Sabuncu, M.R., Liu, H., Schaefer, A., Sun, N., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T.T., 2019. Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion. Cereb. Cortex 29, 2533–2551. https://doi.org/10.1093/cercor/bhy123.

Kong, R., Yang, Q., Gordon, E., Xue, A., Yan, X., Orban, C., Zuo, X.-N., Spreng, N., Ge, T., Holmes, A., Eickhoff, S., Yeo, B.T.T., 2021. Individual-specific areal-level parcellations improve functional connectivity prediction of behavior. Cereb. Cortex 31, 4477–4500. https://doi.org/10.1093/cercor/bhab101.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Kragel, P.A., Han, X., Kraynak, T.E., Gianaros, P.J., Wager, T.D., 2021. Functional MRI can be highly reliable, but it depends on what you measure: a commentary on Elliott et al. (2020). Psychol. Sci. 32, 622–626. https://doi.org/10.1177/0956797621989730.

Larson, R., Ham, M., 1993. Stress and "storm and stress" in early adolescence: the relationship of negative events with dysphoric affect. Dev. Psychol. 29, 130–140.

Laumann, T.O., Gordon, E.M., Adeyemo, B., Snyder, A.Z., Joo, S.J., Chen, M.Y., Gilmore, A.W., McDermott, K.B., Nelson, S.M., Dosenbach, N.U., Schlaggar, B.L., Mumford, J.A., Poldrack, R.A., Petersen, S.E., 2015. Functional system and areal organization of a highly sampled individual human brain. Neuron 87, 657–670.

Laumann, T.O., Snyder, A.Z., Mitra, A., Gordon, E.M., Gratton, C., Adeyemo, B., Gilmore, A.W., Nelson, S.M., Berg, J.J., Greene, D.J., McCarthy, J.E., Tagliazucchi, E., Laufs, H., Schlaggar, B.L., Dosenbach, N.U., Petersen, S.E., 2017. On the stability of BOLD fMRI correlations. Cereb. Cortex 27, 4719–4732.

Lewinsohn, P.M., Gotlib, I.H., Lewinsohn, M., Seeley, J.R., Allan, N.B., 1998. Gender differences in anxiety disorders and anxiety symptoms in adolescents. J. Abnorm. Psychol. 107, 109–117.

Liang, Z., Alberto, M., Martell, C., 2018. Validity of consumer activity wristbands and wearable EEG for measuring overall sleep parameters and sleep structure in freeliving conditions. J. Healthc. Inform. Res. 2, 152–178.

Liu, T.T., 2016. Noise contributions to the fMRI signal: an overview. NeuroImage 143, 141–151. https://doi.org/10.1016/j.neuroimage.2016.09.008.

Maddock, R.J., Garrett, A.S., Buonocore, M.H., 2003. Posterior cingulate cortex activation by emotional words: fMRI evidence from a valence decision task. Hum. Brain Mapp 18, 30–41.

Madhyastha, T., Peverill, M., Koh, N., McCabe, C., Flournoy, J., Mills, K., King, K., Pfeifer, J., McLaughlin, K.A., 2018. Current methods and limitations for longitudinal fMRI analysis across development. Dev. Cogn. Neurosci., Methodol. Challenges Develop. Neuroimag. 33, 118–128. https://doi.org/10.1016/j.dcn.2017.11.006.

Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S., Donohue, M.R., Foran, W., Miller, R.L., Feczko, E., Miranda-Dominguez, O., Graham, A.M., Earl, E.A., Perrone, A.J., Cordova, M., Doyle, O., Moore, L.A., Conan, G., Uriarte, J., Snider, K., Tam, A., Chen, J., Newbold, D.J., Zheng, A., Seider, N.A., Van, A.N., Laumann, T.O., Thompson, W.K., Greene, D.J., Petersen, S.E., Nichols, T.E., Yeo, B.T.T., Barch, D.M., Garavan, H., Luna, B., Fair, D.A., Dosenbach, N.U.F., 2020. Towards reproducible brain-wide association studies. bioRxiv 2020.08.21.257758. https://doi.org/10.1101/2020.08.21.257758.

Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S., Donohue, M.R., Foran, W., Miller, R.L., Hendrickson, T.J., Malone, S.M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A.M., Earl, E.A., Perrone, A.J., Cordova, M., Doyle, O., Moore, L.A., Conan, G.M., Uriarte, J., Snider, K., Lynch, B.J., Wilgenbusch, J.C., Pengo, T., Tam, A., Chen, J., Newbold, D.J., Zheng, A., Seider, N. A., Van, A.N., Metoki, A., Chauvin, R.J., Laumann, T.O., Greene, D.J., Petersen, S.E., Garavan, H., Thompson, W.K., Nichols, T.E., Yeo, B.T.T., Barch, D.M., Luna, B., Fair, D.A., Dosenbach, N.U.F., 2022. Reproducible brain-wide association studies require thousands of individuals. Nature 603, 654–660. https://doi.org/10.1038/s41586-022-04492-9.

Marek, S., Tervo-Clemmens, B., Nielsen, A.N., Wheelock, M.D., Miller, R.L., Laumann, T. O., Earl, E., Foran, W.W., Cordova, M., Doyle, O., Perrone, A., Miranda-Dominguez, O., Feczko, E., Sturgeon, D., Graham, A., Hermosillo, R., Snider, K., Galassi, A., Nagel, B.J., Ewing, S.W.F., Eggebrecht, A.T., Garavan, H., Dale, A.M., Greene, D.J., Barch, D.M., Fair, D.A., Luna, B., Dosenbach, N.U.F., 2019. Identifying reproducible individual differences in childhood functional brain networks: an ABCD study. Dev. Cogn. Neurosci. 40, 100706 https://doi.org/10.1016/j.dcn.2019.100706.

Matta, T.H., Flournoy, J.C., Byrne, M.L., 2018. Making an unknown unknown a known unknown: missing data in longitudinal neuroimaging studies. Dev. Cogn. Neurosci., Methodol. Challeng. Develop. Neuroimag. 33, 83–98. https://doi.org/10.1016/j.dcn.2017.10.001.

McCrory, E.J., De Brito, S.A., Sebastian, C.L., Mechelli, A., Bird, G., Kelly, P.A., Viding, E., 2011. Heightened neural reactivity to threat in child victims of family violence. Curr. Biol. 21, R947–R948.

McLaughlin, K.A., Weissman, D., Bitrán, D., 2019. Childhood adversity and neural development: a systematic review. Annu. Rev. Dev. Psychol. 1, 277–312. https://doi.org/10.1146/annurev-devpsych-121318-084950.

Michl, L.C., McLaughlin, K.A., Shepherd, K., Nolen-Hoeksema, S., 2013. Rumination as a mechanism linking stressful life events to symptoms of depression and anxiety: Longitudinal evidence in early adolescents and adults. J. Abnorm. Psychol. 122, 339–352. https://doi.org/10.1037/a0031994.

Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L., Griffanti, L., Douaud, G., Okell, T.W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P.M., Smith, S.M., 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat. Neurosci. 19, 1523–1536.

Moberly, N.J., Watkins, E.R., 2008. Ruminative self-focus, negative life events, and negative affect. Behav. Res. Ther. 46, 1034–1039.

Monk, C.S., Klein, R.G., Telzer, E.H., Schroth, E.A., Mannuzza, S., Moulton, J.L., Guardino, M., Masten, C.L., McClure-Tone, E.B., Fromm, S., Blair, R.J., Pine, D.S., Ernst, M., 2008. Amygdala and nucleus accumbens activation to emotional facial expressions in children and adolescents at risk for major depression. Am. J. Psychiatry 165, 90–98. https://doi.org/10.1176/appi.ajp.2007.06111917.

Mroczek, D.K., Almeida, D., 2004. The effect of daily stress, personality, and age on daily negative affect. J. Pers. 72, 355–378.

Nook, E.C., Flournoy, J.C., Rodman, A.M., Mair, P., McLaughlin, K.A., 2021. High emotion differentiation buffers against internalizing symptoms following exposure to stressful life events in adolescence: an intensive longitudinal study. Clin. Psychol. Sci. https://doi.org/10.1177/2167702620979786, 2167702620979786.

Northoff, G., Heinzel, A., De Greck, M., Bermpohl, F., Dobrowolny, H., Panksepp, J., 2006. Self-referential processing in our brain—a meta-analysis of imaging studies on the self. NeuroImage 31, 440–457.

Ochsner, K.N., Knierim, K., Ludlow, D.H., Hanelin, J., Ramachandran, T., Glover, G., Mackey, S.C., 2004. Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. J. Cogn. Neurosci. 16, 1746–1772. https://doi.org/10.1162/0898929042947829.

Ordaz, S.J., Foran, W., Velanova, K., Luna, B., 2013. Longitudinal growth curves of brain function underlying inhibitory control through adolescence. J. Neurosci. 33, 18109–18124. https://doi.org/10.1523/JNEUROSCI.1741-13.2013.

O'Toole, M.S., Renna, Megan.E., Elkjær, E., Mikkelsen, M.B., Mennin, D.S., 2020. A systematic review and meta-analysis of the association between complexity of emotion experience and behavioral adaptation. Emot. Rev. 12, 23–38. https://doi.org/10.1177/1754073919876019.

Pashler, H., Harris, C.R., 2012. Is the replicability crisis overblown? Three arguments examined. Perspect. Psychol. Sci. 7, 531–536. https://doi.org/10.1177/1745691612463401.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team, 2020. Nlme: linear and nonlinear mixed effects models.

Poldrack, R.A., 2017. Precision neuroscience: dense sampling of individual brains. Neuron 95, 727–729.

Poldrack, R.A., 2011. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. Neuron 72, 692–697.

Poldrack, R.A., Laumann, T.O., Koyejo, O., Gregory, B., Hover, A., Chen, M.Y., Gorgolewski, K.J., Luci, J., Joo, S.J., Boyd, R.L., Hunicke-Smith, S., Adeyemo, B., Petersen, S.E., Glahn, D.C., McKay, D.R., Curran, J.E., Goring, H.H.H., Carless, M.A., Blangero, J., Dougherty, R., Leemans, A., Handwerker, D.A., Frick, L., Marcotte, E. M., Mumford, J.A., 2015. Long-term neural and physiological phenotyping of a single human. Nat. Commun. 6, 1–15.

Poldrack, R.A., Mumford, J.A., Nichols, T.E., 2011. Handbook of Functional MRI Data Analysis. Cambridge University Press, New York, NY.

Pustejovsky, J., 2019. clubSandwich: Cluster-Robust (Sandwich) variance estimators with small-sample corrections.

Pustejovsky, J.E., Tipton, E., 2018. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. J. Bus. Econ. Stat. 36, 672–683. https://doi.org/10.1080/07350015.2016.1247004.

Python Client. TemplateFlow [WWW Document], n.d. URL https://www.templateflow.org/client/#custom-study-population-specific-templates (accessed 3.2.23).

R Core Team, 2021. R: A language and environment for statistical computing.

Raudenbush, S.W., 2001. Comparing personal trajectories and drawing causal inferences from longitudinal data. Annu. Rev. Psychol. 52, 501–525. https://doi.org/10.1146/annurev.psych.52.1.501.

Revelle, W., Condon, D.M., 2019. Reliability from α to ω: a tutorial. Psychol. Assess., Methodol. Statis. Adv. Clin. Assess. 31, 1395–1411. https://doi.org/10.1037/pas0000754.

Rocke, C., Shu-Chen, L., Smith, J., 2009. Intraindividual variability in positive and negative affect over 45 days: Do older adults fluctuate less than young adults. Psychol. Aging 24, 863–878.

Rodman, A.M., Vidal Bustamante, C.M., Dennison, M.J., Flournoy, J.C., Coppersmith, D. D.L., Nook, E.C., Worthington, S., Mair, P., McLaughlin, K.A., 2021. A year in the social life of a teenager: within-persons fluctuations in stress, phone communication, and anxiety and depression. Clin. Psychol. Sci. https://doi.org/10.1177/21677026219918O4, 2167702621991804.

Rohrer, J.M., Murayama, K., 2023. These are not the effects you are looking for: causality and the within-/between-persons distinction in longitudinal data analysis. Adv. Methods Pract. Psychol. Sci. 6, 25152459221140840 https://doi.org/10.1177/25152459221140842.

Rudolph, K.D., Hammen, C., 1999. Age and gender as determinants of stress exposure, generation, and reactions in youngsters: a transactional perspective. Child Dev. 70, 660–677.

Sabatinelli, D., Fortune, E.E., Li, Q., Siddiqui, A., Krafft, C., Oliver, W.T., Beck, S., Jeffries, J., 2011. Emotional perception: meta-analyses of face and natural scene processing. NeuroImage 54, 2524–2533. https://doi.org/10.1016/j.neuroimage.2010.10.011.

Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T.T., 2018. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cereb. Cortex 28, 3095–3114. https://doi.org/10.1093/cercor/bhx179.

Schiavone, S.R., Quinn, K.A., Vazire, S., 2023. A Consensus-Based Tool for Evaluating Threats to the Validity of Empirical Research. https://doi.org/10.31234/osf.io/fc8v3.

Schiel, J.E., Tamm, S., Holub, F., Petri, R., Dashti, H.S., Domschke, K., Feige, B., Lane, J. M., Riemann, D., Rutter, M.K., Saxena, R., Tahmasian, M., Wang, H., Kyle, S.D., Spiegelhalder, K., 2022. Associations Between Sleep Health and Amygdala Reactivity to Negative Facial Expressions in the UK Biobank Cohort. Biol. Psychiatry, Threat, Stress, and Health 92, 693–700. https://doi.org/10.1016/j.biopsych.2022.05.023.

Sergerie, K., Chochol, C., Armony, J.L., 2008. The role of the amygdala in emotional processing: a quantitative meta-analysis of functional neuroimaging studies. Neurosci. Biobehav. Rev. 32, 811–830.

Shenhav, A., Botvinick, M.M., Cohen, J.D., 2013. The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron 79, 217–240.

Shrout, P., Fleiss, J., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428. https://doi.org/10.1037/0033-2909.86.2.420.

Sliwinski, M.J., Almeida, D., Smyth, J.M., Stawski, R.S., 2009. Intraindividual change and variability in daily stress processes: findings from two measurement-burst diary studies. Psychol. Aging 24, 828–840.

Sliwinski, M.J., Smyth, J.M., Hofer, S.M., Stawski, R.S., 2006. Intraindividual coupling of daily stress and cognition. Psychol. Aging 21, 545–557.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M., Matthews, P.M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23 (Supplement 1), S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051.

Somerville, L.H., Bookheimer, S.Y., Buckner, R.L., Burgess, G.C., Curtiss, S.W., Dapretto, M., Elam, J.S., Gaffrey, M.S., Harms, M.P., Hodge, C., Kandala, S., Kastman, E.K., Nichols, T.E., Schlaggar, B.L., Smith, S.M., Thomas, K.M., Yacoub, E., Van Essen, D.C., Barch, D.M., 2018. The lifespan human connectome project in development: a large-scale study of brain connectivity development in 5–21 year olds. NeuroImage 183, 456–468. https://doi.org/10.1016/j.neuroimage.2018.08.050.

Somerville, L.H., Kim, H.K., 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, 897–903.

Sugiura, M., Shah, N.J., Zilles, K., Fink, G.R., 2005. Cortical representations of personally familiar objects and places: functional organization of the human posterior cingulate cortex. J. Cogn. Neurosci. 17, 183–198.

Swartz, J.R., Knodt, A.R., Radtke, S.R., Hariri, A.R., 2015a. A neural biomarker of psychological vulnerability to future life stress. Neuron 85, 505–511.

Swartz, J.R., Williamson, D.E., Hariri, A.R., 2015b. Developmental change in amygdala reactivity during adolescence: effects of family history of depression and stressful life events. Am. J. Psychiatry 172, 276–283.

Thomas, K.M., Drevets, W.C., Dahl, R.E., Ryan, N.D., Birmaher, B., Eccard, C.H., Axelson, D.A., Whalen, P.J., Casey, B.J., 2001. Amygdala response to fearful faces in anxious and depressed children. Arch. Gen. Psychiatry 58, 1057–1063.

Tottenham, N., Hare, T., Millner, A., Gilhooly, T., Zevin, J.D., Casey, B.J., 2011. Elevated amygdala response to faces following early deprivation. Dev. Sci. 14, 190–204.

Tottenham, N., Phuong, J., Flannery, J., Gabard-Durnam, L., Goff, B., 2013. A negativity bias for ambiguous facial-expression valence during childhood: converging evidence from behavior and facial corrugator muscle responses. Emotion 13, 92–103.

Tottenham, N., Tanaka, J.W., Leon, A.C., McCarry, T., Nurse, M., Hare, T.A., Marcus, D.J., Westerlund, A., Casey, B.J., Nelson, C., 2009. The NimStim set of facial expressions: Judgments from untrained research participants. Psychiatry Res 168, 242–249. https://doi.org/10.1016/j.psychres.2008.05.006.

van der Helm, E., Yao, J., Dutt, S., Rao, V., Saletin, J.M., Walker, M.P., 2011. REM sleep depotentiates amygdala activity to previous emotional experiences. Curr. Biol. 21, 2029–2032. https://doi.org/10.1016/j.cub.2011.10.052.

Vehtari, A., Gabry, J., Yao, Y., Gelman, A., 2018. loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models.

Vermeylen, L., Wisniewski, D., Gonzalez-Garcia, C., Hoofs, V., Notebaert, W., Braem, S., 2020. Shared neural representations of cognitive conflict and negative affect in the medial frontal cortex. J. Neurosci. 40, 8715–8725.

Vidal Bustamante, C.M., Rodman, A.M., Dennison, M.J., Flournoy, J.C., Mair, P., McLaughlin, K.A., 2020. Within-person fluctuations in stressful life events, sleep, and anxiety and depression symptoms during adolescence: a multiwave prospective study. J. Child Psychol. Psychiatry 61, 1116–1125. https://doi.org/10.1111/jcpp.13234.

Wang, L., LaBar, K.S., McCarthy, G., 2006. Mood alters amygdala activation to sad distractors during an attentional task. Biol. Psychiatry 60, 1139–1146.

Wassing, R., Lakbila-Kamal, O., Ramautar, J.R., Stoffers, D., Schalkwijk, F., Van Someren, E.J.W., 2019. Restless REM sleep impedes overnight amygdala adaptation. Curr. Biol. 29, 2351–2358. https://doi.org/10.1016/j.cub.2019.06.034 e4.

Watson, D., Clark, L.A., Tellegan, A., 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. J. Pers. Soc. Psychol. 54, 1063–1070.

Watson, D., Walker, L.M., 1996. The long-term stability and predictive validity of trait measures of affect. J. Pers. Soc. Psychol. 70, 567–577.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. NeuroImage 92, 381–397. https://doi.org/10.1016/j.neuroimage.2014.01.060.

Winkler, A.M., Webster, M.A., Vidaurre, D., Nichols, T.E., Smith, S.M., 2015. Multi-level block permutation. NeuroImage 123, 253–268. https://doi.org/10.1016/j.neuroimage.2015.05.092.

Wood, S.N., 2017. Generalized Additive Models: An Introduction with R, Second Edition. CRC Press.

Xue, A., Kong, R., Yang, Q., Eldaief, M.C., Angeli, P.A., DiNicola, L.M., Braga, R.M., Buckner, R.L., Yeo, B.T.T., 2021. The detailed organization of the human cerebellum estimated by intrinsic functional connectivity within the individual. J. Neurophysiol. 125, 358–384. https://doi.org/10.1152/jn.00561.2020.

Yoo, S.-S., Gujar, N., Hu, P., Jolesz, F.A., Walker, M.P., 2007. The human emotional brain without sleep–a prefrontal amygdala disconnect. Curr. Biol. CB 17, R877–R878. https://doi.org/10.1016/j.cub.2007.08.007.

Zelkowitz, R.L., Cole, D.A., 2016. Measures of emotion reactivity and emotion regulation: convergent and discriminant validity. Personal. Individ. Differ. 102, 123–132. https://doi.org/10.1016/j.paid.2016.06.045.