Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2021 June 28.

Published in final edited form as:

Neuroimage. 2021 May 15; 232: 117872. doi:10.1016/j.neuroimage.2021.117872.

The longitudinal stability of fMRI activation during reward processing in adolescents and young adults

David A.A. Baranger[a,*], **Morgan Lindenmuth**[a], **Melissa Nance**[a], **Amanda E. Guyer**[b,c], **Kate Keenan**[d], **Alison E. Hipwell**[a], **Daniel S. Shaw**[e], **Erika E. Forbes**[a,e]

[a]University of Pittsburgh School of Medicine, Department of Psychiatry, 121 Meyran Avenue, Pittsburgh, PA 15213, United States

[b]Center for Mind and Brain, University of California Davis, Davis, CA, United States

[c]Department of Human Ecology, University of California Davis, Davis, CA, United States

[d]University of Chicago, Department of Psychiatry and Behavioral Neuroscience, Chicago, IL, United States

[e]University of Pittsburgh, Department of Psychology, Pittsburgh, PA, United States

Abstract

Background: The use of functional neuroimaging has been an extremely fruitful avenue for investigating the neural basis of human reward function. This approach has included identification of potential neurobiological mechanisms of psychiatric disease and examination of environmental, experiential, and biological factors that may contribute to disease risk via effects on the reward system. However, a central and largely unexamined assumption of much of this research is that neural reward function is an individual difference characteristic that is relatively stable and trait-like over time.

Methods: In two independent samples of adolescents and young adults studied longitudinally (*Ns* = 145 & 139, 100% female and 100% male, ages 15–21 and 20–22, 2–4 scans and 2 scans respectively), we tested within-person stability of reward-task BOLD activation, with a median of 1 and 2 years between scans. We examined multiple commonly used contrasts of active states and baseline in both the anticipation and feedback phases of a card-guessing reward task. We

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

*Corresponding author. dbaranger@pitt.edu (D.A.A. Baranger).

Authors statement

David AA Baranger : Conceptualization, Formal analysis, Writing - Original Draft, Visualization.
Morgan Lindenmuth : Data Curation, Writing - Review & Editing
Melissa Nance : Data Curation, Writing - Review & Editing
Amanda E. Guyer : Funding acquisition, Project administration, Investigation, Resources, Writing - Review & Editing
Kate Keenan : Funding acquisition, Project administration, Investigation, Resources, Writing - Review & Editing
Alison E Hipwell : Project administration, Investigation, Resources, Writing - Review & Editing
Daniel S Shaw : Funding acquisition, Project administration, Investigation, Resources, Writing - Review & Editing
Erika E Forbes : Supervision, Funding acquisition, Project administration, Investigation, Resources, Writing - Review & Editing

Declaration of Competing Interest

All authors have no conflicts of interest to declare.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.117872.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

examined the effects of cortical parcellation resolution, contrast, network (reward regions and resting-state networks), region-size, and activation strength and variability on the stability of reward-related activation.

Results: In both samples, contrasts of an active state relative to a baseline were more stable (ICC: intra-class correlation; e.g., Win>Baseline; mean ICC = 0.13 – 0.33) than contrasts of two active states (e.g., Win>Loss; mean ICC = 0.048 – 0.05). Additionally, activation in reward regions was less stable than in many non-task networks (e.g., dorsal attention), and activation in regions with greater between-subject variability showed higher stability in both samples.

Conclusions: These results show that some contrasts from functional neuroimaging activation during a card guessing reward task have partially trait-like properties in adolescent and young adult samples over 1–2 years. Notably, results suggest that contrasts intended to map cognitive function and show robust group-level effects (i.e. Win > Loss) may be less effective in studies of individual differences and disease risk. The robustness of group-level activation should be weighed against other factors when selecting regions of interest in individual difference fMRI studies.

Keywords

Reward; Reliability; Longitudinal; Adolescence; Development

1. Introduction

The translational relevance of neural reward processing research is evident from a large and growing literature revealing group differences in regional functional magnetic resonance imaging (fMRI) activation during reward processing across a range of neuropsychiatric conditions, including anxiety, ADHD, depression, addiction, schizophrenia, and autism (Chase et al., 2018; Clements et al., 2018; Forbes et al., 2006; Guyer et al., 2012; Luijten et al., 2017; Ng et al., 2019; Scheres et al., 2007). Building on this work, neural response to reward has been considered a potential biomarker (reflecting the presence of a disorder) or endophenotype (reflecting genetic risk) for these same conditions (Caseras et al., 2013; Dichter, 2012; Grimm et al., 2014; Hasler et al., 2004; Moeller and Paulus, 2018; Pizzagalli, 2014; Rubia, 2018; Sutherland and Stein, 2018). A major implication of this work is that reward-related neural activation may reflect causal neurobiological processes that underlie the emergence of neuropsychiatric disorders. Accordingly, a host of individual difference studies have sought to uncover potential mechanisms that may influence disease risk via their effects on reward processing, such as genetic risk and stress exposure (e.g. Banihashemi et al., 2014; Carey et al., 2017; Casement et al., 2015; Corral-Fríasa et al., 2015; Hanson et al., 2015b, 2015a; Jia et al., 2016; Kumar et al., 2015; Luking et al., 2016; Novick et al., 2018; Romens et al., 2015; Ruggeri et al., 2015). By examining possible causal mechanisms underlying mental illness, findings from this line of research have the potential to have a large impact on the development of novel treatment and prevention measures.

A central assumption of much of this work relating neural response to reward to the presence of, or risk for, a psychiatric disorder is that patterns of neural response to reward, as measured by fMRI, is a stable trait. In this report we distinguish between ‘stability’ and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

'reliability', which are both operationalized as the intra-class correlation ($ICC_{(3,1)}$) between two measurement time points, but differ in the time interval between measurements. We use 'reliability' to refer to measurements separated by days or weeks, and 'stability' to refer to measurements separated by months or years. A measurement of a trait will be both reliable and stable. Only a few studies have examined the stability of neural response to reward stimuli over time. Two prior studies have used different versions of the monetary incentive delay (MID) task, in which participants are told whether the potential outcome of a trial is win, loss, or neutral, and then instructed to respond quickly to a cue (Keren et al., 2018; Wu et al., 2014). Wu et al. and Keren et al. both examined the longitudinal stability (e.g. over 2+ years) of activation using versions of the MID, in adults and adolescents (Ns=14 and 16), respectively. Wu et al. tested the stability of anticipatory activation and observed that the nucleus accumbens and insula were stable only during gain and loss anticipation, respectively (ICCs 0.5–0.7). Keren et al. found stable activation during feedback in the striatum (peak ICC = 0.89) and a stable reward prediction error (RPE) signal in the insula (peak ICC = 0.735). Using a gambling reward Braams et al. examined the nucleus accumbens in the Win>Loss contrast during reward outcome (N = 238, ages 8–26; ICCs=0.219–0.327) (Braams et al., 2015). Korucuoglu et al. examined the longitudinal stability (over 8 months) of activation to the Balloon Analogue Risk Taking (BART) (N = 44, ages 21–24) (Korucuoglu et al., 2020). A range of stabilities were observed, and *a priori* reward regions (left anterior insula and right caudate) were stable during decision making (ICCs of 0.54 and 0.63). This work provides initial evidence that the nucleus accumbens has some stability during reward processing in both adolescents and adults. However, the impact of analytic decisions, such as choice of task phase (i.e. anticipation vs. feedback) and contrast (i.e. Win > Neutral vs. Win > Loss) remains unknown. Furthermore, prior work may suffer from small sample sizes, and the stability of regions outside of the nucleus accumbens —even of regions known to play key roles in reward processing and to be implicated in the etiology of neuropsychiatric disorders (e.g. the orbitofrontal cortex (Ng et al., 2019))— has been largely underexplored. As a result, there is limited evidence that these findings generalize.

Beyond this initial work examining the stability of reward activation, little is known about the factors associated with the reward activation stability. Candidate factors include region of interest (ROI) and region size, activation strength, activation variability, and network membership (e.g. reward, salience, or control). Although prior work has largely focused on regions considered to be part of the canonical reward network, reward tasks invariably elicit activation in other networks - such as in frontoparietal, cingulo-opercular, motor, and visual regions. Furthermore, activation of these regions may be less specific to reward task cues, but they still play important roles in reward function (Haber and Knutson, 2010; Schultz, 2000), and could contribute to differences in reward behavior between diagnostic groups. It is also clear that group-level contrast maps for reward tasks are highly reproducible (e.g., across studies a consistent set of canonical reward regions are reported as activated). (Bartra et al., 2013; Jia et al., 2016; Kampa et al., 2020), suggesting that reward and non-reward regions may differ in their stability. Additionally, evidence suggests that the homogeneity of fMRI activation in a reward task varies as a function of ROI size (Schaefer et al., 2017); given that there are approximately 400 cortical areas (Van Essen et al., 2012), ROI size may

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

influence stability of activation over time. As a whole, more information is needed about the stability of reward activation with regard to these various factors in order to advance longitudinal and clinical neuroscience research.

It is of critical importance to understand the influence of developmental timing on reward stability. Structural and functional brain maturation, including reward function, co-occurs with emergence or worsening of mental health problems during adolescence and young adulthood (Caspi et al., 2020; Foulkes and Blakemore, 2018; Galvan, 2010; Kessler et al., 2005; Kessler and Wang, 2008). Although capturing stability during this period of development is likely to be challenging given individual differences in the pace of maturation, such individual differences would be highly relevant to studies of disease risk, progression, and preventive intervention.

The present study aims to test the stability of individual differences in reward function measured during a card guessing task with monetary reward in two independent, adolescent and young-adult samples (*N*s=145 and 153 respectively). The current research represents the largest study of whole-brain stability in a reward fMRI task to date. We compared the stability of regional activation across commonly used contrasts from this task across the whole brain, and as a function of network identity, ROI size, and activation magnitude and variability. We focused on these aspects in relation to assessing stability of neural response to reward with the goal of informing the optimal analysis of fMRI reward tasks to be used in future individual difference studies. We hypothesized that the stability of reward activation would be higher in young adult than adolescent samples, estimates of activation in canonical reward regions would be more stable over time than in regions within other networks, and that ROI size would be correlated with stability.

2. Methods

Data were drawn from two independent longitudinal neuroimaging studies, the Pitt Mother & Child Project (PMCP) and the Pittsburgh Girls Study – Emotions Substudy (PGS-E), both administered the same reward processing task and were conducted with MRI scanning at the same location. In both studies, all procedures received Institutional Review Board approval at the University of Pittsburgh and all participants provided consent for their participation in the study.

2.1. Pitt mother & child project (PMCP)

The Pitt Mother & Child Project (PMCP) was a longitudinal study of 310 boys and their families residing in low income/resourced environments. Participants were recruited in 1991 and 1992 from Allegheny County Women, Infant and Children (WIC) Nutritional Supplement Clinics when boys were between 6 and 17 months old (Shaw et al., 2012, 2003). Boys and their mothers or primary caregivers were seen almost yearly from ages 1.5 – 22 years in the laboratory and/or home where they completed questionnaires, a psychiatric interview, and, at ages 20 and 22 years, fMRI scanning sessions. Study visits and interviews occurred as close to participants' birthdays as was practically possible. Of the 310 participants, 139 had usable data available for both scans (see Table 1 for demographics). PMCP participants were excluded from the current study if (1) they could not be recruited

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

for an MRI visit (n = 75; i.e., refused or unable to participate), (2) they were ineligible for an MRI scan due to medical or physical reasons (n = 49; i.e., concussion, recent drug use or metal in body), (3) they did not complete the MRI portion of the study (n = 25), (4) they did not correctly perform the task (n = 13; i.e., fell asleep, < 80% accuracy, or did not understand instructions), or (5) the fMRI scan did not pass quality control benchmarks (n = 10; e.g. excessive movement or poor coverage of the nucleus accumbens). The majority of participants were either White American (n = 72, 51.8%) or Black/African American (n = 55, 39.57%), with the remainder identifying as bi-racial (n = 8, 5.76%), or Native Hawaiian, Native American, or Mexican American (n = 4, 0.72%). These racial/ethnic demographics are consistent with city demographics at the time of recruitment. Participants not included in analyses did not differ from those included by age (β=0.21, SE=0.15, t = 1.34, p = 0.17) or by race (χ^2 =0.43, p = 0.52).

2.2. Pittsburgh girls study – emotions sub-study (PGS-E)

Participants were girls and their mothers recruited from the longitudinal Pittsburgh Girls Study (PGS) (Keenan et al., 2010). The PGS sample was formed following an enumeration of households with girls between the ages of 5 and 8 in the city of Pittsburgh. Of the 2990 eligible families, 2450 (85%) were successfully re-contacted and agreed to participate in a prospective study. A subset of PGS participants was recruited to the PGS Emotions Sub-study (PGS-E) a study of precursors to depression (N = 232 (Keenan et al., 2008)). From ages 15–21, participants were invited to complete four annual study visits and fMRI scanning sessions (see Table 1 for demographics). Of the potential n = 928 scans (4 per participant), scans were unavailable or excluded from analyses because (1) participants withdrew or could not be scheduled (n = 228 scans across n = 121 participants), (2) participants were ineligible for an MRI scan due to medical or physical reasons (n = 80 scans, n = 55 participants), (3) participants refused or were unable to complete the MRI portion of the study (n = 69,n = 48 participants), (4) participants did not correctly perform the reward task or scans did not pass MRI quality control benchmarks (n = 86, n = 65 participants), or (5) because the aforementioned reasons resulted in only one usable MRI session (n = 38 participants). The final sample consisted of 439 scans from n = 145 participants, 49 participants had four scans, 51 had three scans, and 45 had two scans (Figure S1). The average time between scans was 1.27 years (SD=0.5), which ranged from 0.44 to 4.18 years. The majority of participants were Black/African American (n = 101, 69.6%) or White American (n = 35, 24%), with the remainder identifying as multi-racial (n = 9, 4.8%). Participants not included in analyses did not differ from those included by age (β=0.22, SE=0.12, t = 1.88, p = 0.06) and were more likely to be white (χ^2 =4.35, p = 0.037; 38% of participants excluded vs. 24% of participants included were white).

2.3. Monetary reward and loss fMRI task

Participants in both samples completed the same version of an 8-minute slow event-related card-guessing task involving anticipation and receipt of monetary reward (Figure S2) (Forbes et al., 2010). In each trial (20 s), participants were shown a card with a question mark, told the possible value of the card was 1–9, and asked to guess whether the card's value was lower or higher than 5 (4 s), rendering a card value of 5 neutral. Participants were then cued to the trial type (6 s) – possible-win or possible-loss (i.e., the anticipation phase).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Participants were told the “correct” answer (500 ms) and then whether they had a positive (gain money), negative (lose money), or neutral (no-change) outcome (500 ms; i.e., the outcome phase). Each trial ended with a cross-hair that was presented during a 9 second inter-trial interval. Participants completed 24 trials, with a balanced number of trial types (i.e., 12 possible-win and 12 possible-loss, 6 win-outcome, 6 loss-oucome, 12 neutral-outcome). Trials were presented in a pseudorandom order with predetermined outcomes. Participants received $10 after completing the task.

2.4. Neuroimaging data collection and preprocessing

Data collection: All participants were scanned using the acquisition protocols on the same Siemens 3T Trio scanner, which was not upgraded or modified over the course of either study. BOLD functional images were acquired with a gradient echo planar imaging (EPI) sequence and covered 39 axial slices (3.1 mm wide) beginning at the cerebral vertex and extending across the entire cerebrum and the majority of the cerebellum (TR/RE=2000/25 ms, flip angle=90°, field of view =20 cm, matrix=64×64). A reference EPI scan was acquired before fMRI data collection, which was visually inspected for artifacts (e.g., ghosting) and for adequate signal across the entire volume. A 160-slice, high-resolution, sagittally acquired T1-weighted anatomical image (MPRAGE) was collected for co-registration and normalization of functional images (TR/TE = 2300/2.98 ms, field of view = 20 cm, matrix = 256 × 240).

Preprocessing and within-person analysis: Preprocessing for both samples was completed using Statistical Parametric Mapping software (SPM8; http://www.fil.ion.ucl.ac.uk/spm). Structural images for each participant were auto-segmented, and functional images were realigned to correct for head motion, registered to the segmented structural data, spatially normalized into standard stereotaxic space (Montreal Neurological Institute template) using a 12-parameter affine model, and smoothed with a 6 mm full-width at half-maximum Gaussian filter. Voxel-wise signal was ratio-normalized to the whole-brain global mean. The Artifact Detection Toolbox (ART; http://www.nitrc.org/projects/artifact_detect/) software was used to detect functional volumes with movement > 3 SD from the subject’s mean, > 0.5 mm scan-to-scan translation, or > 0.01° of scan-to-scan rotation. Preprocessed data were inspected to ensure that all participants had fewer than 25% of volumes with excessive movement detected by ART, good scan quality, and ventral striatum coverage of at least 80%. Temporal censoring based on ART output was used to remove motion artifacts in first-level analyses.

The two projects (PMCP and PGS-E) used slightly different time intervals within the course of the task to define the different phases of reward processing (e.g. anticipation and outcome). These differences have been consistent across the publications within each study (Casement et al., 2016, 2015, 2014; Hasler et al., 2017; Morgan et al., 2014; Romens et al., 2015). Thus, to maintain consistency with prior work and to compare the effects of differences in task phase definition within each study, no adjustments were made in the present analyses. In the PMCP, reward anticipation was defined as the 6 s when the symbol indicating trial-type was displayed, and reward outcome was defined as the 1 second of outcome and the first 6 s of the inter-trial interval. In the PGS-E, reward anticipation

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

included the 6 s when the symbol indicating trial-type was displayed as well as the subsequent 2 s, to account for the delay in hemodynamic response relative to neural activity and to capture as much of the reward anticipation response as possible while avoiding substantial overlap with BOLD response to reward outcome events. However, as the PGS-E definition of reward anticipation included the first 2 s of outcome (i.e., prior to the onset of the hemodynamic response to the outcome cue), the outcome phase was not modeled with its own regressor in the GLM. We note that analyses of the PMCP did model the outcome phase with its own regressor, and findings from both samples reach similar conclusions (see Results). As such, we do not think this decision is driving our findings. For both studies, baseline was defined as the last 3 s of the inter-trial interval (during which a fixation was presented). This period was modeled with a dedicated regressor to compute a relatively neutral estimate of the average baseline activation across all trials, uncontaminated by hemodynamic responses to the task stimuli. The Decision phase of the task was not modeled with a regressor, as it is too short (4 s) to capture the hemodynamic response to the decision cue without being influenced by the response to the anticipation cue. First-level general linear models (GLMs) were used to calculate images for all contrasts. In both samples these included: (1) win anticipation > loss anticipation, (2) win anticipation > baseline, and (3) loss anticipation > baseline. In the PMCP four additional contrasts were calculated: (4) win outcome > loss outcome, (5) win outcome > baseline, (6) loss outcome > baseline, (7) win outcome > neutral outcome (e.g. no-win and no-loss), and (8) loss outcome > neutral outcome (Supplemental Figures 3-5).

Cortical regions of interest (ROIs) were defined using the Schaefer atlas (Schaefer et al., 2017), a recent cortical parcellation derived from resting-state fMRI data. One strength of this atlas, for the present analyses, is that parcellations at different resolutions (e.g., 100 – 1000 regions) were identified, using identical methodology and data. Thus, the use of this atlas permits examination of the influence of ROI size, while holding potential confounds constant, including parcellation method (automatic vs. manual) and source data (histological vs. DTI vs. resting-state fMRI). To allow the broadest possible definition of regional response, all of the Schaefer atlas parcellations – 100, 200, 300, 400, 500, 600, 700, 800, and 1000 ROIs – were used. Subcortical ROIs were identified using the Harvard-Oxford subcortical atlas (Frazier et al., 2005) using a threshold of 25% probability, and voxels present across multiple ROIs were discarded. All cortical and subcortical ROIs were included in analyses. Subject-level average percent signal-change was extracted from each ROI from every contrast in both samples, using the Marsbar toolbox for SPM (Brett et al., 2002).

2.5. Statistical analyses

2.5.1. Stability—The longitudinal test-retest reliability ($ICC_{(3,1)}$), which we refer to here as ‘stability’ (Becht and Mills, 2020), of the activation of each region for each contrast was estimated using the R software package ‘RptR’ (Stoffel et al., 2017), in which a mixed-effect linear model was fit to all observations from all participants, with a random intercept for each participant. Stability was then estimated as the ratio of the variance of within-person means over the sum of the group-level and residual variance. That is, reliability in the context of a longitudinal study is an estimate of the stability of individual differences over

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

time (Revelle and Condon, 2018). As it is a ratio, this measure is bounded by [0, 1], though in cases where it is extremely small (i.e. $<2.2 \times 10^{-16}$) it is reported as 0. Higher values, closer to 1, indicate that the measurement is more consistent over time. Models estimated the stability of each ROI in each sample, across all ROIs in every Schaefer parcellation (4600 ROIs total across 9 parcellations) and the Harvard-Oxford atlas (14 ROIs), for all contrasts (8 contrasts in the PMCP, 3 in the PGS-E). Participant race was added as a fixed effect, as were age and age^2, as reward-related fMRI activation is known to vary as a function of age (Braams et al., 2015; Lamm et al., 2014). Regional activation was winsorized to 3 standard deviations prior to reliability estimation, to reduce the influence of outliers.

Post-hoc analyses additionally examined the stability of *ranked* activation in the PMCP and PGS-E (Supplemental methods), as it is possible that group-level changes in mean activation, beyond what is captured by participant age, could bias stability estimates. The stability of ranked activation was strongly correlated with primary stability estimates across all parcellations, contrasts, and regions (PMCP: $r = 0.965$, $p<2.2 \times 10^{-16}$; PGS: $r = 0.60$, $p<2.2 \times 10^{-16}$). Results of these analyses are reported in the Supplement methods, as they do not meaningfully differ from the results and conclusions of the primary analyses.

2.5.2. Analyses—Linear mixed-effect models were used to test associations of parcellation, contrast, and ROI, entered as crossed random effects with random intercepts, with ROI stability. Stability estimates were winsorized to three standard deviations for these analyses, to reduce the influence of outliers. Variance explained by random effects was assessed using repeatability (R; i.e., the ratio of variance explained over total variance) calculated using the Rptr package (Stoffel et al., 2017). Once it emerged that parcellations did not differ in stability, and that the Win > Baseline contrast during the Anticipation phase had the highest stability (see Results), subsequent results considered only this contrast, using the $n = 400$ Schaefer parcellation, given estimates of approximately 400 cortical regions (Van Essen et al., 2012), with the addition of subcortical ROIs. Analyses then examined the correlation between ROI stability and network identity (entered as dummy-variables), ROI size, average activation, and the variation of activation in both the PMCP and PGS-E samples. Average activation and the variation in activation of each ROI were estimated as the intercept and standard-error of the intercept in a linear mixed effect model with no fixed-effects, and participant as a random intercept. All analyses were conducted in R (R Core Team, 2014). As the smallest number that many R packages will report is 2.2×10^{-16} – p-values lower than this value are reported at $p<2.2 \times 10^{-16}$.

2.5.3. ROI network assignment—The Schaefer parcellations include an assignment of each cortical ROI to one of the seven canonical non-overlapping resting state networks: visual, somatomotor, dorsal attention, ventral attention, limbic, frontoparietal, and default mode (Yeo et al., 2011). However, these networks do not include subcortical regions, and there is no one network specific to the cognitive demands of the reward task used in the present analyses. Thus, an eighth ‘reward’ network was defined using Neurosynth (Yarkoni et al., 2011). Neurosynth is a platform that generates meta-analyses, using reported loci from published fMRI studies, thus identifying the regions of the brain most likely to be reported in studies on a given topic. We used the Neurosynth meta-analysis of reward-associated

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

keywords, including “reward”, “outcome”, “anticipation”, and “monetary” (e.g., topic 7 from the v5 50-topic solution (Poldrack et al., 2012), which identifies voxels (p<0.01 false-discovery rate corrected) more likely to be reported in reward studies (n = 1218) than non-reward studies (n = 13,153) (Supplemental Figure 6), https://neurosynth.org/analyses/topics/v5-topics-50/7. ROIs’ network-assignment was changed from Schaefer-assigned networks to the reward network, based on the overlap between voxels in the Neurosynth meta-analysis and each ROI. A t-statistic was calculated based on the distribution of this overlap and regions that showed a significant overlap (p<0.05) were assigned to the reward network. The reward network consisted of 22 regions, including the bilateral nucleus accumbens, caudate, putamen, pallidum, amygdala, and thalamus. Cortical regions included ROIs in the bilateral ventromedial prefrontal cortex and orbitofrontal cortex. The hippocampus was the only subcortical region not assigned to the reward network by the Neurosynth meta-analysis – it was assigned to the limbic network, as it has long been recognized as a central component of that system (Morgane et al., 2005). In analyses where networks were compared, the limbic network was set as the reference network, as it was observed to have the lowest stability (see Results).

3. Results

3.1. Parcellation resolution does not impact stability

Mixed effect models, in which parcellation was treated as a random intercept, revealed that stability did not differ across parcellations in the PMCP sample (R = 0, p = 1), but did in PGS-E sample. This minimal effect of parcellation (R = 0.004, p = 0.001) was driven by slightly lower stability in the 100- and 200-region parcellations (Supplemental Figure 7). The addition of a second random intercept for task contrasts improved model fit and explained a significant amount of the variance of ROI stability (PMCP: R = 0.64, $p < 2.2 \times 10^{-16}$; PGS-E: R = 0.32, $p < 2.2 \times 10^{-16}$). There was no evidence for an interaction between parcellation and contrasts, as the addition of an interaction term did not improve model fit (PMCP: $\chi^2_{(1)} = 0$, p = 1; PGS-E: $\chi^2_{(1)} = 1.9$, p = 0.166) (Supplemental Figure 8). The addition of a random intercept for each ROI explained a significant amount of the total variation in stability (PMCP: R = 0.11 $p < 2.2 \times 10^{-16}$; PGS-E: R = 0.266, $p < 2.2 \times 10^{-16}$), indicating that the stability of individual ROIs was somewhat consistent across contrasts. The stability of every ROI across all contrasts in both samples is provided in the Supplemental Data File.

3.2. Reward anticipation and contrasts relative to baseline are more stable than loss contrasts or contrasts between active conditions

Given the evidence that stability did not vary as a function of parcellation resolution, analyses were then restricted to the Schaefer et al. parcellation with 400 parcels, based on estimates that there are approximately 400 cortical areas (Van Essen et al., 2012); subcortical regions were included in all subsequent analyses. In the PMCP sample, stability of average whole-brain activation was highest during reward anticipation in the Win > Baseline contrast (Fig. 1A; $\bar{x}$ = 0.327, SD = 0.091). Win contrasts were associated with higher activation stability than Loss contrasts (β = 0.045, SE = 0.003, t = 15.71, $p < 2.2 \times 10^{-16}$), and contrasts relative to Baseline were associated with higher stability (β = 0.203, SE = 0.003, t

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

= 70.77, $p < 2.2 \times 10^{-16}$) (Fig. 1A, Table S1). The feedback phase was associated with lower activation stability than the anticipation phase ($\beta = -0.015$, $SE = 0.003$, $t = -5.29$, $p = 1.31 \times 10^{-7}$). In the PGS-E sample, which only examined the anticipation phase, activation estimates from the Win and Loss contrasts, each relative to Baseline, were similarly more stable than those from the contrast between these two active conditions ($\beta = 0.064$, SE = 0.003, $t = 19.54$, $p < 2.2 \times 10^{-16}$) (Fig. 1B, Table S1). However, during anticipation the Win > Baseline contrast was slightly less stable, on average, than Loss > Baseline ($\beta = -0.008$, $SE = 0.003$, $t = -2.65$, $p = 0.008$; $\bar{x} = 0.115$, $SD = 0.07$).

3.3. Stability is lower in the limbic and reward networks than other cortical networks

Differences between the average stability of networks during the anticipation of rewards (Anticipation > Baseline) was examined next. In the PMCP sample, results from an ANOVA showed a significant effect of network ($F_{7, 406} = 15.5$, $p < 2.2 \times 10^{-16}$, adjusted R2 = 0.197). Relative to the limbic network, the reward network did not significantly differ, whereas for all others (i.e., frontoparietal, dorsal attention, salience, default mode, visual, and somatomotor) activation estimates were more stable (Fig. 2A&C, Table S2 and S3). In the PGS-E sample, there was similarly a significant effect of network ($F_{7, 403} = 11.47$, $p = 2.67 \times 10^{-13}$, adjusted $R^2 = 0.152$). The limbic network was also the least stable in the PGS-E (Fig. 2B and D, Table S2), and all networks were more stable than it, except for the reward network, which was not significantly different from the limbic network (Table S3).

3.4. Regions with higher between-subject variability are more stable

Average between-subject activation and activation variability – the intercept and SE of the intercept from mixed effect models predicting the activation of each region – as well as region-size, were then added to the linear model predicting stability. In the PMCP, these terms significantly improved model fit ($F_{3, 403} = 43.38$, $p < 2.2 \times 10^{-16}$, change in adjusted $R^2 = 0.19$). Average activation was associated with lower stability ($\beta = -0.09$, standardized $\beta = -0.154$, $SE = 0.0168$, $t = -5.55$, $p = 5.16 \times 10^{-8}$), whereas activation variability was associated with greater stability ($\beta = 4.78$, standardized $\beta = 0.400$, $SE = 0.517$, $t = 9.26$, $p < 2.2 \times 10^{-16}$) (Fig. 3). ROI size (# of voxels) was nominally associated with stability in the full model ($\beta = 5.6 \times 10^{-5}$, standardized $\beta = 0.160$, $SE = 2.5 \times 10^{-5}$, $t = 2.32$, $p = 0.026$). In the PGS-E sample, these terms similarly significantly improved model fit ($F_{3,400} = 53.89$, $p < 2.2 \times 10^{-16}$, delta adjusted $R^2 = 0.24$). As in the PMCP sample, activation variation in the PGS-E sample was associated with greater stability ($\beta = 5.745$, standardized $\beta = 0.52$, $SE = 0.455$, $t = 12.617$, $p < 2.2 \times 10^{-16}$). Average activation was not associated with stability ($\beta = 0.002$, standardized $\beta = -0.030$, $SE = 0.001$, $t = 1.37$, $p = 0.17$), and ROI size was similarly nominally associated with stability ($\beta = 4.2 \times 10^{-5}$, standardized $\beta = 0.137$, $SE = 1.8 \times 10^{-5}$, $t = 2.235$, $p = 0.026$) (Fig. 3). In both samples, activation variation was significantly associated with network (PMCP: $F_{7, 406} = 10.89$, $p = 1.3 \times 10^{-12}$, adjusted $R^2 = 0.144$; PGS-E: $F_{7, 403} = 12.09$, $p = 4.9 \times 10^{-14}$, adjusted $R^2 = 0.159$), wherein variation was lowest in the limbic and reward networks, and significantly higher in every other network (Supplemental Table 4). The replicable and strong association of activation variation with stability indicates that regions with greater between-person variation also tend to be the most stable over time.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

To further assess the agreement in results between the two samples, the linear models correlating stability of activation estimates during the anticipation phase for the Win > Baseline contrast were used to predict stability across samples. Models included network identity, ROI size, and the mean and standard deviation of activation. Stability in the PMCP sample was predicted using results from the PGS-E, and conversely stability in the PGS-E sample was predicted using results from the PMCP. The predicted and true stabilities were significantly correlated in both samples (PMCP: $r = 0.38$, $r^2=0.14$, $p = 1.9 \times 10^{-15}$; PGS-E: $r = 0.35$, $r^2=0.12$, $p = 1.7 \times 10^{-13}$). In contrast, the measured stability of each individual region in the PMCP was only weakly correlated with the stability of each region in the PGS-E ($r = 0.158$, $p = 0.0013$). Thus, while the stability of individual regions differs between the samples, both samples agree on several factors that influence stability.

4. Discussion

Our goal in the present study was to contribute data on the stability of reward function as measured by fMRI. This is one of the first studies to examine stability in relatively large samples of males and females, over a 1–2 year interval, and during the transition from adolescence to early adulthood. Our findings replicate results from the few existing studies; some reward contrasts, particularly contrasts of an active state against a baseline, are relatively stable over 1–2 years during this developmental period. In particular, while the magnitude of activation in many regions during the *anticipation* of reward was stable, few regions were stable when contrasting *win* relative to *loss*, a widely used contrast in studies of reward processing. Surprisingly, although we observed lower stability for the core reward-processing network, some of the highest stability estimates were found outside of that network. In addition, regions with the greatest between-person variability tended to show the highest level of stability.

4.1. Contrasts of active conditions are less stable

One of the most striking observations from the present study is that contrasts of two active conditions (e.g. Win > Loss) were less stable than contrasts of a task condition to a fixation baseline. This was true in two independent samples, and in the PMCP sample the pattern was evident for both anticipation and outcome contrasts. Thus, we suggest that contrasts of two active task conditions, particularly of the reward task used here and similar tasks, may not be suitable as biomarkers for psychopathologies and may not reflect genetic or environmental risk for psychopathology, as they demonstrate poor stability. Thus, between-person differences in such contrasts may not be attributable to the presence of underlying traits.

The limited stability of Win > Loss activation in the present study adds to a growing body of work highlighting concerns over the reliability of fMRI task measures in general (Elliott et al., 2020; Fliessbach et al., 2010; Infantolino et al., 2018). The common practice of defining ‘activation’ as the difference in the average percent signal change between two different task states may contribute to the low reliability and stability reported for fMRI tasks (Luking et al., 2017). fMRI task activations computed this way are difference-scores; the reliability of a difference-score is partially a function of the correlation between the items being subtracted,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and is lower for items that are more strongly correlated (Thomas and Zumbo, 2012). When two conditions, such as the anticipation of uncertain monetary *gain* or *loss*, are strongly correlated, the resulting difference score will have poor reliability.

Indeed, we note that studies of short-term reliability of reward activation yield stronger estimates when the contrast included a neutral comparison. Several studies have reported a moderate-to-good short-term (days or weeks) reliability of activation of the ventral striatum (ICC = 0.5 - 0.8) when examining the anticipation of monetary gain relative to a neutral condition where neither gain or loss was possible (Holiga et al., 2018; Plichta et al., 2012; Schlagenhauf et al., 2008). The neutral condition used in these studies differs from the neutral condition in the present study, a no-change outcome (for the PMCP sample only), which might be interpreted as a positive or negative outcome, depending on whether the trial involved not-losing or not-winning, respectively. Similarly, Wu et al. observed some evidence for improved stability when considering the raw average fMRI signal, relative to a contrast of an active and neutral condition (Wu et al., 2014). In contrast, other studies have examined the short-term reliability of gain vs. loss contrasts, observing lower reliabilities in the striatum (ICC = −0.13 – 0.45) (Elliott et al., 2020; Fliessbach et al., 2010; Li et al., 2020). Luking et al. compared the split-half reliability of gain>baseline and gain>loss activation during reward outcome, finding nearly double the reliability in gain>baseline in the ventral striatum (Luking et al., 2017). These results are congruent with our own results, suggesting that fMRI contrasts will be less stable if the cognitive processes underlying trial-types are too similar. However, the extent to which task design, versus contrast of choice, influences the reliability of fMRI signals remains an underexplored area.

4.2. Additional factors influencing stability

Beyond comparing contrasts, other factors that influence longitudinal stability were identified and found to be consistent across the samples used in this study, particularly the observation that greater between-subject variability of activation was associated with increased stability, and that ‘task specific’ ROIs tended not to be the most stable. The observation that more variable between-subject activation—in other words, regions that exhibited a greater range of individual differences—was predictive of greater longitudinal stability is not unique to this study. This pattern has been reported previously in different tasks (Fröhner et al., 2019) and modalities (resting-state (Pannunzi et al., 2017)), and has been discussed at length in the psychometric literature (e.g., the reverse, that effects with less between-subject variation are less reliable is sometimes referred to as the ‘reliability paradox’ (Hedge et al., 2018)). This observation in the present study may very well be a consequence of the computation of fMRI activation as a difference-score.

Interestingly, and contrary to our hypothesis, we observed that activation measured from task-specific ROIs tends to be less stable than that from non-task ROIs. In the present study, ‘task-specific’ ROIs were located primarily in the subcortex and ventral surface of the cortex, regions which are well-known to suffer from increased artifact susceptibility (Merboldt et al., 2000; Wiggins et al., 2009), which may reduce stability. Additionally, while most studies that have examined the reliability or stability of reward task activation restricted their analyses to *a priori* reward-network ROIs, some looked beyond these structures. These

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

studies have found that it is not uncommon for other regions of the brain to exhibit greater reliability and stability than the regions targeted by a task (i.e., regions in the default mode network have been frequently observed to be highly reliable) (Elliott et al., 2020; Fröhner et al., 2019; Holiga et al., 2018; Keren et al., 2018; Vetter et al., 2017). These tasks are designed to elicit strong within-subject effects in targeted regions at a single time-point. This results in lower between-subject variance in targeted regions (Hedge et al., 2018), leading to reduced reliability. Non-targeted regions, which perform computations not directly manipulated by the task, thus exhibit greater between-subject variability, and are hence more reliable. Indeed, we observed lower between-subject variability, as well as lower stability, in task-specific ROIs.

Several factors were not replicably predictive of stability, including activation strength and region size. The association between activation strength and stability was inconsistent across samples, with a negative association in the PMCP and weak positive association in the PGS-E. The prior literature is also inconsistent, as studies and meta-analyses have reported both positive and null correlations between activation and reliability (Bennett and Miller, 2010; Elliott et al., 2020; Fliessbach et al., 2010; Korucuoglu et al., 2020; Li et al., 2020; Plichta et al., 2012). This aligns with work demonstrating that robust group-level between-condition differences do not imply that a measurement is reliable (Infantolino et al., 2018; Matheson, 2019).

The present analyses used a set of cortical parcellations (Schaefer et al., 2017) that ranged in the number of regions, from 100 to 1000. Doing so allowed us to explore the possible effects of ROI size and parcellation resolution, independent of potential confounding effects of parcellation method (e.g., DTI vs histology vs resting state fMRI). These effects were inconsistent across studies and in opposing directions. The parcellations with the largest ROIs (the 100 and 200 parcellations) were less stable in the PGS-E, but there was no effect on stability in the PMCP. In contrast, ROI size was very weakly associated with increased stability in the Schaefer-400 parcellation both samples. These results, which run contrary to our initial hypothesis, suggest that the choice of larger or smaller ROIs will not impact the sensitivity of individual-difference analyses, and that ROI choice should be guided by other considerations. For instance, smaller ROIs allow for the examination of more fine-grained effects on activation. However, this issue needs to be balanced with the consequent cost of a larger correction for multiple comparisons, which results in reduced power.

4.3. Interpretation of stability in developing samples

Although we found that contrasts of neural activation to active condition relative to baseline were more stable than contrasts between different task conditions, our observed stabilities were not in the range that is typically considered “good” for test-retest reliability (e.g., ICC > 0.7) (Cicchetti, 1994). However, this standard comes from studies of inter-rater and short-term reliability, and there are no widely-accepted standards for stability. When reliability is used as a statistic reflecting measurement error it is assumed that error is the primary source of any difference between two observations. Measurement error will clearly not be the only source of difference for many measurements that are separated by years or in samples that are developing rapidly (Streiner and Norman, 2008). Studies of psychiatric diagnoses and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

psychopathology in adolescents and adults have reported a wide range of stabilities (Blázquez et al., 2019; Olino et al., 2018; Pettit et al., 2005; Shankman et al., 2017), as have studies of cognition in adolescents (Taylor et al., 2020), suggesting that a stability estimate greater than 0.7 is not necessary for a measurement to reflect meaningful individual differences. Prior work has observed that fMRI activation can be within this range when certain contrasts are used (see the Discussion above), yet the same benchmarks are likely inappropriate for stability across several years. Indeed, meta-analyses have found that the reliability of both fMRI activation (Bennett and Miller, 2010) and resting-state correlations (Noble et al., 2019) decrease with longer intervals between scans (but see (Elliott et al., 2020)), and it is generally expected, irrespective of measurement type, that stability will decrease with longer intervals between assessments (Streiner and Norman, 2008). As an increasing number of longitudinal studies of brain function are conducted, it will be critical to report the stability of these measures, so that appropriate benchmarks can be determined (Herting et al., 2018).

Based on the large body of literature detailing the influence of development, aging, and life experience on brain function, it would be surprising if the longitudinal stability of neural activation estimates were as high as its short-term stability, especially during periods of known developmental change in the circuitry of interest (Braams et al., 2015; Galvan et al., 2006). Indeed, emerging work has suggested that the longitudinal stability of activation in feedback-learning and rule-learning fMRI tasks is highest in adult populations (Koolschijn et al., 2011; Peters et al., 2016). In the current study, the age of both samples provided the developmental window of late adolescence to early adulthood (15–22 years, inclusive) within which to measure brain function. This approach may have also contributed to the relatively lower observed stability. However, we note that the observed range of stabilities in the present analyses are comparable to longitudinal stabilities observed in adolescent samples performing other tasks, including cognitive control (Vetter et al., 2017), performance monitoring (Koolschijn et al., 2011), rule-learning (Peters et al., 2016), and emotional face monitoring (van den Bulk et al., 2013). Examining the influence of development on the stability of fMRI activation will be an important direction for future work.

We also note an alternative interpretation of the present results. While some prior studies on the short-term reliability (e.g., 2 weeks) of reward activation have reported reliabilities above our current results (ICC = 0.5 - 0.8) (Holiga et al., 2018; Plichta et al., 2012; Schlagenhauf et al., 2008), others report short-term reliabilities in the same range as those reported in the present study, which had a long test-retest interval (ICC = −0.13 – 0.45) (Elliott et al., 2020; Fliessbach et al., 2010; Li et al., 2020). As discussed in Section 4.1, we attribute these differences to methodological differences between the studies. However, it is quite possible that studies reporting higher levels of reliability are outliers, and that the reliability of reward activation is in the lower range regardless of the duration between scans. This could, for instance, be due to technical aspects of fMRI data collection (see Section 4.4, below), limiting the precision of the fMRI measurements. Studies with estimates of both short-term reliability and long-term stability are needed to reconcile these differing interpretations. Hence, it is clear that work improving the precision of fMRI will continue to be extremely valuable.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

While Win>Baseline activation is partially trait-like during adolescence and young adulthood, the observed stabilities of reward activation (mean ICC: 0.13 – 0.33) indicate that one measurement is not strongly predictive of a second measurement a year or two later. Whether or not the stability of reward function is adequate will depend on the research context (Streiner and Norman, 2008). We recommend that researchers planning future longitudinal studies examining within-subject trajectories of reward activation account for this variance in their power calculations, as failing to do so may result in severely underpowered samples (Guo et al., 2013; Raudenbush and Xiao-Feng, 2001). We do not provide specific recommendations here as power in a longitudinal analysis is affected by a myriad of factors and design choices beyond just sample size (e.g. number of observations, time between observations, and sample missingness).

4.4. Factors that may limit the maximally observable stability

Many factors beyond interscan interval and development are likely influence the stability of neural activation as measured by fMRI tasks. Evidence suggests that activation during cognitive and affective tasks is sensitive to time-varying states, including mood, sleep, and recent stress (Nikolova and Hariri, 2012; Hasler et al., 2012; Baranger et al., 2016, 2017), which may place a ceiling on the maximum stability that is achievable with task-based fMRI. This concern is in addition to the wide array of technical aspects of fMRI data collection that will influence the signal-to-noise ratio, including task design, scanner manufacturer, acquisition protocol, ambient temperature, and head motion (Elliott et al., 2020; Greve et al., 2011; Karch et al., 2019; Petersen and Dubis, 2012; Power et al., 2014). Indeed, the task used here is a variant of a widely used design (Delgado et al., 2000), and contains all of the hallmarks of a reward processing task (Richards et al., 2013), most notably the receipt of a reward at the end of a trial, as well as trials that begin with a cue indicating reward potential. While prior work has suggested that the activation and reliability of reward tasks is largely consistent across different designs (Fliessbach et al., 2010; Sescousse et al., 2013), it should be noted that our task and analyses include features that, while not uncommon, may have influenced stability measures, including reward outcomes that are not contingent on the participant's behavior, a small number of trials, an inter-trial-interval that is not jittered, and analyses conducted with older software versions (i.e. SPM8 vs SPM12). Indeed, the number of trials in the task used is smaller than current recommendations (Murphy and Garavan, 2005). As the task is an older slow event-related design, fewer trials were used to reduce participant burden. An important limitation of the current study is that the task included a low number of trials for generating the most accurate (Li et al., 2020) and replicable (Nee, 2019) estimates. Nonetheless, the available datasets analyzed in the current study provided an important preliminary opportunity to address questions about the long-term stability of reward function. Furthermore, using the same task design as in past work facilitates interpretation of the current findings relative to the literature on reward function using this task. Thus, despite the low number of task trials, the current results are useful in generating hypotheses about the long-term stability of reward function and point to where replication with newer designs will be necessary.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

4.5. Comparisons between samples

One of the major strengths of the present report is the consistency of results across two samples that differed in several respects. Participants in the PMCP were all male, while participants in the PGS-E were all female. The ages of the two samples only partially overlapped – PGS-E participants (ages 15–21) were younger than PMCP participants (ages 20–22). However, there were some differences in the results from the two samples, particularly the lower overall stability in the PGS-E. While this observation is consistent with our hypothesis that stability would be lower in adolescent samples, we note that there are several additional factors that may have driven this difference, including differences in sample demographics and study methodology (e,g. participant sex or recruitment approach).

It is well documented that reward-related neural and behavioral processes undergo developmental changes into the 20 s (Casey et al., 2008), and participants in both samples were likely undergoing development in reward and cognitive control circuitry, both of which mature during this period. We hypothesized that stability would be lower in the adolescent PGS-E sample, as greater age-related change occurs during this time period. The effects of development are not uniform - individuals' trajectories of change occur at different times and rates (Chahal et al., 2018) – which would be expected to reduce the stability of measurements taken prior to, or during, the developmental period.

The PMCP sample was entirely male, whereas the PGS-E was entirely female. Prior reports have observed greater activation during reward processing in males than females (Alarcón et al., 2017; Spreckelmeyer et al., 2009), which could be attributable to differential effects of sex hormones on reward sensitivity (Dreher et al., 2007; Harden et al., 2018). The samples also differed by socioeconomic status. All the participants in the PMCP came from families that were receiving economic benefits for low-income families (WIC) at the time of recruitment. The PGS over-sampled homes in low-resourced neighborhoods in the enumeration, resulting in approximately 40% of participants coming from families that received public assistance (Keenan et al., 2010). As a result, monetary rewards may not have had the same saliency across samples, which could produce differences in reward activation (Zink et al., 2004), resulting in stability differences.

The PMCP and PGS-E also took different approaches to scheduling participants. PMCP participants completed fMRI sessions as close to their 20th and 22nd birthdays as was practical, resulting in very little variation in the time between measurements (Supplemental Figure 1). The PGS-E took the more common approach of allowing flexibility in scheduling participants for their annual visit. Additionally, as the PGS-E collected fMRI data over a longer period, participants could miss one follow-up, but return for the subsequent visit, which was not possible in the PMCP, which included only two waves of fMRI data collection. These differences in study design resulted in much more variation in the time between measurements in the PGS-E, which may have contributed to the lower stability seen in the PGS-E sample. The additional waves of data collection in the PGS-E may have also contributed to the lower stability by virtue of reduced task novelty, though whether detectable habituation effects are present in fMRI measurements separated by years remains a largely unexplored topic (Telzer et al., 2018). Finally, the two studies differed in their definition of the anticipation phase of the task. In the PMCP, this phase ended once the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

feedback cue was presented, whereas in the PGS-E it was extended to include the feedback cue presentation, to account for the delay in the hemodynamic response to the anticipation cue. It is possible that activation during the extended anticipation phase may have been contaminated with activation reflecting subsequent processing.

4.6. Future directions

The present study compared the stability of commonly used contrasts derived from an fMRI reward task, yet further work is needed to establish benchmarks for stability. Meeting these benchmarks may require the development of new tasks or paradigms, such as naturalistic imaging (Gruskin et al., 2019), or analyzing current tasks in different ways. For example, some reports suggest that correlations between trial-wise participant behavior (e.g., subjective value) and neural activation have increased reliability, relative to contrasts of activation in response to task conditions (Keren et al., 2018), but see also Fliessbach (Fliessbach et al., 2010). One intriguing recent report suggests that the low reliability of classic cognitive behavioral tasks can be resolved with hierarchical models that account for variability in trial-level behavioral responses, as op-posed to simply averaging behavior across trials (Haines et al., 2020). A recent application of a similar approach to neuroimaging data suggests that it results in increased power (Chen et al., 2020). Related work has found that latent variable modeling can be used to improve the ability of fMRI analyses to detect individual differences (Cooper et al., 2019). Another promising approach is the use of multivariate models, in which MRI data are used to create a model predicting brain-states for independent samples, which show evidence of having greater power than mass univariate tests (Marek et al., 2020). Combining these approaches may prove to be a fruitful area of future research, particularly as they could be used to improve analyses of existing task fMRI data sets.

Conclusions

The findings from the present study suggest that in late adolescence through early adulthood, neural activation derived from a reward fMRI task has partially trait-like properties over 1–2 years, with activation estimates in a contrast of two active conditions less stable than in a contrast of an active condition to a baseline. This difference likely accounts for some of the discrepancies in prior reports on the reliability of fMRI activation. In addition, regions with greater between-subject variability and regions within non-task networks exhibited activation with greater stability, indicating that the robustness of group-level activation is not a sufficient sole criterion for selecting brain regions to use in individual difference studies. Whether these observed stabilities are adequate will depend on researcher’s questions and study designs. Further work is needed to establish benchmarks for the stability of fMRI tasks, and to develop tasks and analytic pipelines optimized for improving stability of neural activation measured with fMRI tasks.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Acknowledgments

The authors would like to thank the participants from both studies. This work was supported by grants from the National Institutes of Mental Health, including DA026222 (Shaw, Forbes), MH093605 (Guyer, Keenan, Forbes), and MH018951 (Baranger).

Data availability

Summary data used in these analyses are provided in the Supplemental Data file. Formal data sharing proposal and agreement forms for the Pitt Mother & Child Project can be requested from senior authors Dr. Shaw (danielshaw@pitt.edu) or Dr. Forbes (forbese@upmc.edu). Formal data sharing proposal and agreement forms for the Pitsburgh Girls Study - Emotions Substudy can be requested from senior authors Dr. Keenan (kekeenan@uchicago.edu) or Dr. Hipwell (hipwae@upmc.edu).

References

Alarcón G, Cservenka A, Nagel BJ, 2017. Adolescent neural response to reward is related to participant sex and task motivation. Brain Cogn. 111, 51–62. doi:10.1016/j.bandc.2016.10.003. [PubMed: 27816780]

Banihashemi L, Sheu LK, Midei AJ, Gianaros PJ, 2014. Cumulative stress in childhood is associated with blunted reward-related brain activity in adulthood, 1–29.

Baranger DAA, Ifrah C, Prather AA, Carey CE, Corral-Frías NS, Drabant Conley E, Hariri AR, Bogdan R, 2016. PER1 rs3027172 genotype interacts with early life stress to predict problematic alcohol use, but not reward-related ventral striatum activity. Front. Psychol 7, 1–10. doi:10.3389/fpsyg.2016.00464. [PubMed: 26858668]

Baranger DAA, Margolis S, Hariri AR, Bogdan R, 2017. An earlier time of scan is associated with greater threat-related amygdala reactivity. Soc. Cogn. Affect. Neurosci 12, 1272–1283. doi:10.1093/scan/nsx057 [PubMed: 28379578]

Bartra O, McGuire JT, Kable JW, 2013. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. Neuroimage 76, 412–427. doi:10.1016/j.neuroimage.2013.02.063. [PubMed: 23507394]

Becht AI, Mills KL, 2020. Modeling individual differences in brain development. Biol. Psychiatry, Converg. Heterogen. Psychopathol 88, 63–69. doi:10.1016/j.biopsych.2020.01.027.

Bennett CM, Miller MB, 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci 1191, 133–155. doi:10.1111/j.1749-6632.2010.05446.x. [PubMed: 20392279]

Blázquez A, Ortiz AE, Castro-Fornieles J, Morer A, Baeza I, Martínez E, Lázaro L, 2019. Five-year diagnostic stability among adolescents in an inpatient psychiatric unit. Compr. Psychiatry 89, 33–39. doi:10.1016/j.comppsych.2018.11.011. [PubMed: 30583125]

Braams BR, van Duijvenvoorde ACK, Peper JS, Crone EA, 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. doi:10.1523/JNEUROSCI.4764-14.2015. [PubMed: 25948271]

Brett M, Anton J-L, Valabregue R, Poline J-B, 2002. Region of interest analysis using an SPM toolbox.

Carey CE, Knodt AR, Conley ED, Hariri AR, Bogdan R, 2017. Reward-related ventral striatum activity links polygenic risk for attention-deficit/hyperactivity disorder to problematic alcohol use in young adulthood. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 2, 180–187. doi:10.1016/j.bpsc.2016.10.003. [PubMed: 28825048]

Casement MD, Guyer AE, Hipwell AE, McAloon RL, Hoffmann AM, Keenan KE, Forbes EE, 2014. Girls' challenging social experiences in early adolescence predict neural response to rewards and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

depressive symptoms. Dev. Cogn. Neurosci., Developmental Social and Affective Neuroscience 8, 18–27. doi:10.1016/j.dcn.2013.12.003.

Casement MD, Keenan KE, Hipwell AE, Guyer AE, Forbes EE, 2016. Neural reward processing mediates the relationship between insomnia symptoms and depression in adolescence. Sleep 39, 439–447. doi:10.5665/sleep.5460. [PubMed: 26350468]

Casement MD, Shaw DS, Sitnick SL, Musselman SC, Forbes EE, 2015. Life stress in adolescence predicts early adult reward-related brain function and alcohol dependence. Soc. Cogn. Affect. Neurosci 10, 416–423. doi:10.1093/scan/nsu061 [PubMed: 24795442]

Caseras X, Lawrence NS, Murphy K, Wise RG, Phillips ML, 2013. Ventral striatum activity in response to reward: differences between bipolar I and II disorders. Am. J. Psychiatry 170, 533–541. doi:10.1176/appi.ajp.2012.12020169. [PubMed: 23558337]

Casey BJ, Getz S, Galvan A, 2008. The adolescent brain. Dev. Rev., Curr. Directions Risk Decis. Mak 28, 62–77. doi:10.1016/j.dr.2007.08.003.

Caspi A, Houts RM, Ambler A, Danese A, Elliott ML, Hariri A, Harrington H, Hogan S, Poulton R, Ramrakha S, Rasmussen LJH, Reuben A, Richmond-Rakerd L, Sugden K, Wertz J, Williams BS, Moffitt TE, 2020. Longitudinal assessment of mental health disorders and comorbidities across 4 decades among participants in the dunedin birth cohort study. JAMA Netw. Open 3, e203221. doi:10.1001/jamanetworkopen.2020.3221. [PubMed: 32315069]

Chahal R, Vilgis V, Grimm KJ, Hipwell AE, Forbes EE, Keenan K, Guyer AE, 2018. Girls' pubertal development is associated with white matter microstructure in late adolescence. Neuroimage 181, 659–669. doi:10.1016/j.neuroimage.2018.07.050. [PubMed: 30056197]

Chase HW, Loriemi P, Wensing T, Eickhoff SB, Nickl-Jockschat T, 2018. Meta-analytic evidence for altered mesolimbic responses to reward in schizophrenia. Hum. Brain Mapp 39, 2917–2928. doi:10.1002/hbm.24049. [PubMed: 29573046]

Chen G, Padmala S, Chen Y, Taylor PA, Cox RW, Pessoa L, 2020. To pool or not to pool: can we ignore cross-trial variability in FMRI? bioRxiv 2020.05.19.102111. 10.1101/2020.05.19.102111

Cicchetti DV, 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess 6, 284–290. doi:10.1037/1040-3590.6.4.284.

Clements CC, Zoltowski AR, Yankowitz LD, Yerys BE, Schultz RT, Herrington JD, 2018. Evaluation of the social motivation hypothesis of autism. JAMA Psychiatry 75, 797–808. doi:10.1001/jamapsychiatry.2018.1100. [PubMed: 29898209]

Cooper SR, Jackson JJ, Barch DM, Braver TS, 2019. Neuroimaging of individual differences: a latent variable modeling perspective. Neurosci. Biobehav. Rev 98, 29–46. doi:10.1016/j.neubiorev.2018.12.022. [PubMed: 30611798]

Corral-Fríasa NS, Nikolovaa YS, Michalskia LJ, Baranger DAA, Hariria AR, Bogdan R, Corral-Frías NS, Nikolova YS, Michalski LJLJ, Baranger DAA, Hariri AR, Bogdan R, 2015. Stress-related anhedonia is associated with ventral striatum reactivity to reward and transdiagnostic psychiatric symptomatology. Psychol. Med 45, 2605–2617. doi:10.1017/S0033291715000525. [PubMed: 25853627]

Delgado MR, Nystrom LE, Fissell C, Noll DC, Fiez J.a, 2000. Tracking the hemodynamic responses to reward and punishment in the striatum. J. Neurophysiol 84, 3072–3077. [PubMed: 11110834]

Dichter GS, 2012. Functional magnetic resonance imaging of autism spectrum disorders. Dialogues Clin. Neurosci 14, 319–351. [PubMed: 23226956]

Dreher J-C, Schmidt PJ, Kohn P, Furman D, Rubinow D, Berman KF, 2007. Menstrual cycle phase modulates reward-related neural function in women. Proc. Natl. Acad. Sci 104, 2465–2470. doi:10.1073/pnas.0605569104. [PubMed: 17267613]

Elliott ML, Knodt AR, Ireland D, Morris ML, Poulton R, Ramrakha S, Sison ML, Moffitt TE, Caspi A, Hariri AR, 2020. What i s the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. Psychol. Sci 31, 792–806. doi:10.1177/0956797620916786. [PubMed: 32489141]

Fliessbach K, Rohe T, Linder NS, Trautner P, Elger CE, Weber B, 2010. Retest reliability of reward-related BOLD signals. Neuroimage 50, 1168–1176. doi:10.1016/j.neuroimage.2010.01.036. [PubMed: 20083206]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Forbes E, Olino T, Ryan N, 2010. Reward-related brain function as a predictor of treatment response in adolescents with major depressive disorder. Cogn. Affect. Behav. Neurosci 10, 107–118. doi:10.3758/CABN.10.1.107.Reward-Related. [PubMed: 20233959]

Forbes EE, Christopher May J, Siegle GJ, Ladouceur CD, Ryan ND, Carter CS, Birmaher B, Axelson DA, Dahl RE, 2006. Reward-related decision-making in pediatric major depressive disorder: an fMRI study. J. Child Psychol. Psychiatry 47, 1031–1040. doi:10.1111/j.1469-7610.2006.01673.x. [PubMed: 17073982]

Foulkes L, Blakemore SJ, 2018. Studying individual differences in human adolescent brain development. Nat. Neurosci 21, 315–323. doi:10.1038/s41593-018-0078-4. [PubMed: 29403031]

Frazier JA, Chiu S, Breeze JL, Makris N, Lange N, Kennedy DN, Herbert MR, Bent EK, Koneru VK, Dieterich ME, Hodge SM, Rauch SL, Grant PE, Cohen BM, Seidman LJ, Caviness VS, Biederman J, 2005. Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. Am. J. Psychiatry 162, 1256–1265. doi:10.1176/appi.ajp.162.7.1256. [PubMed: 15994707]

Fröhner JH, Teckentrup V, Smolka MN, Kroemer NB, 2019. Addressing the reliability fallacy in fMRI: similar group effects may arise from unreliable individual effects. Neuroimage 195, 174–189. doi:10.1016/j.neuroimage.2019.03.053. [PubMed: 30930312]

Galvan, 2010. Adolescent development of the reward system. Front. Hum. Neurosci 4, 1–9. doi:10.3389/neuro.09.006.2010. [PubMed: 20204154]

Galvan A, Hare TA, Parra CE, Penn J, Voss H, Glover G, Casey BJ, 2006. Earlier development of the accumbens relative to orbitofrontal cortex might underlie risk-taking behavior in adolescents. J. Neurosci 26, 6885–6892. doi:10.1523/JNEUROSCI.1062-06.2006. [PubMed: 16793895]

Greve DN, Mueller BA, Liu T, Turner JA, Voyvodic J, Yetter E, Diaz M, Mc-Carthy G, Wallace S, Roach BJ, Ford JM, Mathalon DH, Calhoun VD, Wible CG, Potkin SG, Glover G, 2011. A novel method for quantifying scanner instability in fMRI. Magn. Reson. Med 65, 1053–1061. doi:10.1002/mrm.22691. [PubMed: 21413069]

Grimm O, Heinz A, Walter H, Kirsch P, Erk S, Haddad L, Plichta MM, Romanczuk-Seiferth N, Pöhland L, Mohnke S, Mühleisen TW, Mattheisen M, Witt SH, Schäfer A, Cichon S, Nöthen M, Rietschel M, Tost H, Meyer-Lindenberg A, 2014. Striatal response to reward anticipation: evidence for a systems-level intermediate phenotype for schizophrenia. JAMA Psychiatry 71, 531–539. doi:10.1001/jamapsychiatry.2014.9. [PubMed: 24622944]

Gruskin DC, Rosenberg MD, Holmes AJ, 2019. Relationships between depressive symptoms and brain responses during emotional movie viewing emerge in adolescence. Neuroimage 116217. doi:10.1016/j.neuroimage.2019.116217. [PubMed: 31628982]

Guo Y, Logan HL, Glueck DH, Müller KE, 2013. Selecting a sample size for studies with repeated measures. BMC Med. Res. Methodol 13, 100. doi:10.1186/1471-2288-13-100. [PubMed: 23902644]

Guyer AE, Choate VR, Detloff A, Benson B, Nelson EE, Perez-Edgar K, Fox NA, Pine DS, Ernst M, 2012. Striatal functional alteration during incentive anticipation in pediatric anxiety disorders. Am. J. Psychiatry 169, 205–212. [PubMed: 22423352]

Haber SN, Knutson B, 2010. The reward circuit: linking primate anatomy and human imaging. Neuropsychopharmacol. Off. Publ. Am. Coll. Neuropsychopharmacol 35, 4–26. doi:10.1038/npp.2009.129.

Haines N, Kvam PD, Irving L, Tucker Smith C, Beauchaine TP, Pitt MA, Ahn W, Turner BM, 2020. Learning from the Reliability Paradox: How Theoretically In-formed Generative Models Can Advance the Social, Behavioral, and Brain Sciences. PsyArXiv doi:10.31234/osf.io/xr7y3.

Hanson JL, Albert D, Iselin AMR, Carré JM, Dodge KA, Hariri AR, 2015a. Cumulative stress in childhood is associated with blunted reward-related brain activity in adulthood. Soc. Cogn. Affect. Neurosci 11, 405–412. doi:10.1093/scan/nsv124. [PubMed: 26443679]

Hanson JL, Hariri AR, Williamson DE, 2015b. Blunted ventral striatum development in adolescence reflects emotional neglect and predicts depressive symptoms. Biol. Psychiatry 78, 598–605. doi:10.1016/j.biopsych.2015.05.010. [PubMed: 26092778]

Harden KP, Mann FD, Grotzinger AD, Patterson MW, Steinberg L, Tackett JL, Tucker-Drob EM, 2018. Developmental differences in reward sensitivity and sensation seeking in adolescence:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

testing sex-specific associations with gonadal hormones and pubertal development. J. Pers. Soc. Psychol 115, 161–178. doi:10.1037/pspp0000172. [PubMed: 29094961]

Hasler BP, Casement MD, Sitnick SL, Shaw DS, Forbes EE, 2017. Eveningness among late adolescent males predicts neural reactivity to reward and alcohol dependence 2 years later. Behav. Brain Res 327, 112–120. doi:10.1016/j.bbr.2017.02.024. [PubMed: 28254633]

Hasler BP, Dahl RE, Holm SM, Jakubcak JL, Ryan ND, Silk JS, Phillips ML, Forbes EE, 2012. Weekend–weekday advances in sleep timing are associated with altered reward-related brain function in healthy adolescents. Biol. Psychol 91, 334–341. doi:10.1016/j.biopsycho.2012.08.008. [PubMed: 22960270]

Hasler G, Drevets WC, Manji HK, Charney DS, 2004. Discovering endophenotypes for major depression. Neuropsychopharmacology 29, 1765–1781. doi:10.1038/sj.npp.1300506. [PubMed: 15213704]

Hedge C, Powell G, Sumner P, 2018. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Behav. Res. Methods 50, 1166–1186. doi:10.3758/s13428-017-0935-1. [PubMed: 28726177]

Herting MM, Gautam P, Chen Z, Mezher A, Vetter NC, 2018. Test-retest reliability of longitudinal task-based fMRI: implications for developmental studies. Dev. Cogn. Neurosci 33, 17–26. doi:10.1016/j.dcn.2017.07.001. [PubMed: 29158072]

Holiga Š, Sambataro F, Luzy C, Greig G, Sarkar N, Renken RJ, Marsman J-BC, Schobel SA, Bertolino A, Dukart J, 2018. Test-retest reliability of task-based and resting-state blood oxygen level dependence and cerebral blood flow measures. PLoS ONE 13, e0206583. doi:10.1371/journal.pone.0206583. [PubMed: 30408072]

Infantolino ZP, Luking KR, Sauder CL, Curtin JJ, Hajcak G, 2018. Robust is not necessarily reliable: from within-subjects fMRI contrasts to between-subjects comparisons. Neuroimage 173, 146–152. doi:10.1016/j.neuroimage.2018.02.024. [PubMed: 29458188]

Jia T, Macare C, Desrivières S, Gonzalez DA, Tao C, Ji X, Ruggeri B, Nees F, Banaschewski T, Barker GJ, Bokde ALW, Bromberg U, Büchel C, Conrod PJ, Dove R, Frouin V, Gallinat J, Garavan H, Gowland PA, Heinz A, Ittermann B, Lathrop M, Lemaitre H, Martinot J-L, Paus T, Pausova Z, Poline J-B, Rietschel M, Robbins T, Smolka MN, Müller CP, Feng J, Rothenfluh A, Flor H, Schumann G, 2016. Neural basis of reward anticipation and its genetic determinants doi:10.1073/pnas.1503252113.

Kampa M, Schick A, Sebastian A, Wessa M, Tüscher O, Kalisch R, Yuen K, 2020. Replication of fMRI group activations in the neuroimaging battery for the Mainz Resilience Project (MARP). Neuroimage, 204 doi:10.1016/j.neuroimage.2019.116223.

Karch JD, Filevich E, Wenger E, Lisofsky N, Becker M, Butler O, Mårtensson J, Lindenberger U, Brandmaier AM, Kühn S, 2019. Identifying predictors of within-person variance in MRI-based brain volume estimates. Neuroimage 200, 575–589. doi:10.1016/j.neuroimage.2019.05.030. [PubMed: 31108215]

Keenan K, Hipwell A, Chung T, Stepp S, Stouthamer-Loeber M, Loeber R, McTigue K, 2010. The Pittsburgh Girls Study: overview and initial findings. J. Clin. Child Adolesc. Psychol. Off. J. Soc. Clin. Child Adolesc. Psychol. Am. Psychol. Assoc. Div 53 (39), 506–521. doi:10.1080/15374416.2010.486320.

Keenan K, Hipwell A, Feng X, Babinski D, Hinze A, Rischall M, Henneberger A, 2008. Subthreshold symptoms of depression in preadolescent girls are stable and predictive of depressive disorders. J. Am. Acad. Child Adolesc. Psychiatry 47, 1433–1442. doi:10.1097/CHI.0b013e3181886eab. [PubMed: 19034189]

Keren H, Chen G, Benson B, Ernst M, Leibenluft E, Fox NA, Pine DS, Stringaris A, 2018. Is the encoding of reward prediction error reliable during development? Neuroimage 178, 266–276. doi:10.1016/j.neuroimage.2018.05.039. [PubMed: 29777827]

Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE, 2005. Life-time prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. Arch. Gen. Psychiatry 62, 593–602. doi:10.1001/arch-psyc.62.6.593. [PubMed: 15939837]

Kessler RC, Wang PS, 2008. The descriptive epidemiology of commonly occurring mental disorders in the United States. Annu. Rev. Public Health 29, 115–129. doi:10.1146/annurev.publhealth.29.020907.090847. [PubMed: 18348707]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Koolschijn PCMP, Schel MA, Rooij M.de, Rombouts SARB, Crone EA, 2011. A three-year longitudinal functional magnetic resonance imaging study of performance monitoring and test-retest reliability from childhood to early adulthood. J. Neurosci 31, 4204–4212. doi:10.1523/JNEUROSCI.6415-10.2011. [PubMed: 21411661]

Korucuoglu O, Harms MP, Astafiev SV, Kennedy JT, Golosheykin S, Barch DM, Anokhin AP, 2020. Test-retest reliability of fMRI-measured brain activity during decision making under risk. Neuroimage 214, 116759. doi:10.1016/j.neuroimage.2020.116759. [PubMed: 32205253]

Kumar P, Slavich GM, Berghorst LH, Treadway MT, Brooks NH, Dutra SJ, Greve DN, O'Donovan A, Bleil ME, Maninger N, Pizzagalli DA, 2015. Perceived life stress exposure modulates reward-related medial prefrontal cortex responses to acute stress in depression. J. Affect. Disord 180, 104–111. [PubMed: 25898329]

Lamm C, Benson BE, Guyer AE, Perez-Edgar K, Fox NA, Pine DS, Ernst M, 2014. Longitudinal study of striatal activation to reward and loss anticipation from mid-adolescence into late adolescence/early adulthood. Brain Cogn., Special Issue Reward Regul. Process. Adolesc 89, 51–60. doi:10.1016/j.bandc.2013.12.003.

Li X, Pan Y, Fang Z, Lei H, Zhang X, Shi H, Ma N, Raine P, Wetherill R, Kim JJ, Wan Y, Rao H, 2020. Test-retest reliability of brain responses to risk-taking during the balloon analogue risk task. Neuroimage 209, 116495. doi:10.1016/j.neuroimage.2019.116495. [PubMed: 31887425]

Luijten M, Schellekens AF, Kühn S, MacHielse MWJ, Sescousse G, 2017. Disruption of reward processing in addiction: an image-based meta-analysis of functional magnetic resonance imaging studies. JAMA Psychiatry 74, 387–398. doi:10.1001/jamapsychiatry.2016.3084. [PubMed: 28146248]

Luking KR, Nelson BD, Infantolino ZP, Sauder CL, Hajcak G, 2017. Internal consistency of functional magnetic resonance imaging and electroencephalography measures of reward in late childhood and early adolescence. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 2, 289–297. doi:10.1016/j.bpsc.2016.12.004. [PubMed: 29057369]

Luking KR, Pagliaccio D, Luby JL, Barch DM, 2016. Depression risk predicts blunted neural responses to gains and enhanced responses to losses in healthy children. J. Am. Acad. Child Adolesc. Psychiatry 55, 328–337. doi:10.1016/j.jaac.2016.01.007. [PubMed: 27015724]

Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, Donohue MR, Foran W, Miller RL, Feczko E, Miranda-Dominguez O, Graham AM, Earl EA, Perrone AJ, Cordova M, Doyle O, Moore LA, Conan G, Uriarte J, Snider K, Tam A, Chen J, Newbold DJ, Zheng A, Seider NA, Van AN, Laumann TO, Thompson WK, Greene DJ, Petersen SE, Nichols TE, Yeo BTT, Barch DM, Garavan H, Luna B, Fair DA, Dosenbach NUF, 2020. Towards reproducible brain-wide association studies. bioRxiv 2020.08.21.257758. 10.1101/2020.08.21.257758

Matheson GJ, 2019. We need to talk about reliability: making better use of test-retest studies for study design and interpretation. PeerJ 7. doi:10.7717/peerj.6918 .

Merboldt KD, Finsterbusch J, Frahm J, 2000. Reducing inhomogeneity artifacts in functional MRI of human brain activation-thin sections vs gradient compensation. J. Magn. Reson. San Diego Calif 145, 184–191. doi:10.1006/jmre.2000.2105, 1997.

Moeller SJ, Paulus MP, 2018. Toward biomarkers of the addicted human brain: using neuroimaging to predict relapse and sustained abstinence in substance use disorder. Prog. Neuropsychopharmacol. Biol. Psychiatry 80, 143–154. doi:10.1016/j.pnpbp.2017.03.003. [PubMed: 28322982]

Morgan JK, Shaw DS, Forbes EE, 2014. Maternal depression and warmth during childhood predict age 20 neural response to reward. J. Am. Acad. Child Adolesc. Psychiatry 53, 108–117. doi:10.1016/j.jaac.2013.10.003, e1. [PubMed: 24342390]

Morgane PJ, Galler JR, Mokler DJ, 2005. A review of systems and networks of the limbic forebrain/limbic midbrain. Prog. Neurobiol 75, 143–160. doi:10.1016/j.pneurobio.2005.01.001. [PubMed: 15784304]

Murphy K, Garavan H, 2005. Deriving the optimal number of events for an event-related fMRI study based on the spatial extent of activation. Neuroimage 27, 771–777. doi:10.1016/j.neuroimage.2005.05.007. [PubMed: 15961321]

Nee DE, 2019. fMRI replicability depends upon sufficient individual-level data. Commun. Biol 2, 1–4. doi:10.1038/s42003-019-0378-6. [PubMed: 30740537]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Ng TH, Alloy LB, Smith DV, 2019. Meta-analysis of reward processing in major depressive disorder reveals distinct abnormalities within the reward circuit. Transl. Psychiatry 9, 293. doi:10.1038/s41398-019-0644-x. [PubMed: 31712555]

Nikolova YS, Hariri AR, 2012. Neural responses to threat and reward interact to predict stress-related problem drinking: a novel protective role of the amygdala. Biol. Mood Anxiety Disord 2, 19. doi:10.1186/2045-5380-2-19. [PubMed: 23151390]

Noble S, Scheinost D, Constable RT, 2019. A decade of test-retest reliability of functional connectivity: a systematic review and meta-analysis. Neuroimage, 203 doi:10.1016/j.neuroimage.2019.116157.

Novick AM, Levandowski ML, Laumann LE, Philip NS, Price LH, Tyrka AR, 2018. The effects of early life stress on reward processing. J. Psychiatr. Res 101, 80–103. doi:10.1016/j.jpsychires.2018.02.002. [PubMed: 29567510]

Olino TM, Bufferd SJ, Dougherty LR, Dyson MW, Carlson GA, Klein DN, 2018. The development of latent dimensions of psychopathology across early childhood: stability of dimensions and moderators of change. J. Abnorm. Child Psychol 46, 1373–1383. doi:10.1007/s10802-018-0398-6. [PubMed: 29359267]

Pannunzi M, Hindriks R, Bettinardi RG, Wenger E, Lisofsky N, Martensson J, Butler O, Filevich E, Becker M, Lochstet M, Kuhn S, Deco G, 2017. Resting-state fMRI correlations: from link-wise unreliability to whole brain stability. Neuroimage 157, 250–262. doi:10.1016/j.neuroimage.2017.06.006. [PubMed: 28599964]

Peters S, Van Duijvenvoorde ACK, Koolschijn PCMP, Crone EA, 2016. Longitudinal development of frontoparietal activity during feedback learning: contributions of age, performance, working memory and cortical thickness. Dev. Cogn. Neurosci 19, 211–222. doi:10.1016/j.dcn.2016.04.004. [PubMed: 27104668]

Petersen SE, Dubis JW, 2012. The mixed block/event-related design. Neuroimage 62, 1177–1184. doi:10.1016/j.neuroimage.2011.09.084. [PubMed: 22008373]

Pettit JW, Morgan S, Paukert AL, 2005. The stability of axis I diagnoses in youth across multiple psychiatric hospitalizations. Child Psychiatry Hum. Dev 36, 53–71. doi:10.1007/s10578-004-3493-6. [PubMed: 16049644]

Pizzagalli DA, 2014. Depression, stress, and anhedonia: toward a synthesis and integrated model. Annu. Rev. Clin. Psychol 10, 393–423. doi:10.1146/annurev-clinpsy-050212-185606. [PubMed: 24471371]

Plichta MM, Schwarz AJ, Grimm O, Morgen K, Mier D, Haddad L, Gerdes ABM, Sauer C, Tost H, Esslinger C, Colman P, Wilson F, Kirsch P, Meyer-Lindenberg A, 2012. Test–retest reliability of evoked BOLD signals from a cognitive–emotive fMRI test battery. Neuroimage 60, 1746–1758. doi:10.1016/j.neuroimage.2012.01.129. [PubMed: 22330316]

Poldrack RA, Mumford JA, Schonberg T, Kalar D, Barman B, Yarkoni T, 2012. Discovering relations between mind, brain, and mental disorders using topic mapping. PLOS Comput. Biol 8, e1002707. doi:10.1371/journal.pcbi.1002707. [PubMed: 23071428]

Power JD, Mitra A, Laumann TO, Snyder AZ, Schlaggar BL, Petersen SE, 2014. Methods to detect, characterize, and remove motion artifact in resting state fMRI. Neuroimage 84, 320–341. doi:10.1016/j.neuroimage.2013.08.048. [PubMed: 23994314]

R Core Team, 2014. R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Raudenbush SW Xiao-Feng L, 2001. Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. Psychol. Methods 6, 387–401. [PubMed: 11778679]

Revelle W, Condon DM, 2018. Reliability from alpha to omega: a tutorial (preprint). PsyArXiv. 10.31234/osf.io/2y3w9

Richards JM, Plate RC, Ernst M, 2013. A systematic review of fMRI reward paradigms used in studies of adolescents vs. adults: the impact of task design and implications for understanding neurodevelopment. Neurosci. Biobehav. Rev 37, 976–991. doi:10.1016/j.neubiorev.2013.03.004. [PubMed: 23518270]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Romens SE, Casement MD, McAloon R, Keenan K, Hipwell AE, Guyer AE, Forbes EE, 2015. Adolescent girls' neural response to reward mediates the relation between childhood financial disadvantage and depression. J. Child Psychol. Psychiatry 56, 1177–1184. doi:10.1111/jcpp.12410. [PubMed: 25846746]

Rubia K, 2018. Cognitive neuroscience of attention deficit hyperactivity disorder (ADHD) and its clinical translation. Front. Hum. Neurosci 12. doi:10.3389/fn-hum.2018.00100.

Ruggeri B, Nymberg C, Vuoksimaa E, Lourdusamy A, Wong CP, Carvalho FM, Jia T, Cattrell A, Macare C, Banaschewski T, Barker GJ, Bokde ALW, Bromberg U, Büchel C, Conrod PJ, Fauth-Bühler M, Flor H, Frouin V, Gallinat J, Garavan H, Gowland P, Heinz A, Ittermann B, Martinot J-L, Nees F, Pausova Z, Paus T, Rietschel M, Robbins T, Smolka MN, Spanagel R, Bakalkin G, Mill J, Sommer WH, Rose RJ, Yan J, Aliev F, Dick D, Kaprio J, Desrivières S, Schumann G, Consortium, IMAGEN, 2015. Association of protein phosphatase PPM1G with alcohol use disorder and brain activity during behavioral control in a genome-wide methylation analysis. Am. J. Psychiatry 172, 543–552. doi:10.1176/appi.ajp.2014.14030382. [PubMed: 25982659]

Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo X, Holmes AJ, Eickhoff SB, Yeo BTT, 2017. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cereb. Cortex 1–20. doi:10.1093/cercor/bhx179. [PubMed: 28365777]

Scheres A, Milham MP, Knutson B, Castellanos FX, 2007. Ventral striatal hyporesponsiveness during reward anticipation in attention-deficit/hyperactivity disorder. Biol. Psychiatry 61, 720–724. doi:10.1016/j.biopsych.2006.04.042. [PubMed: 16950228]

Schlagenhauf F, Juckel G, Koslowski M, Kahnt T, Knutson B, Dembler T, Kienast T, Gallinat J, Wrase J, Heinz A, 2008. Reward system activation in schizophrenic patients switched from typical neuroleptics to olanzapine. Psychopharmacology (Berl.) 196, 673–684. doi:10.1007/s00213-007-1016-4. [PubMed: 18097655]

Schultz W, 2000. Multiple reward signals in the brain. Nat. Rev. Neurosci 1, 199–207. doi:10.1038/35044563. [PubMed: 11257908]

Sescousse G, Caldú X, Segura B, Dreher J-C, 2013. Processing of primary and secondary rewards: a quantitative meta-analysis and review of human functional neuroimaging studies. Neurosci. Biobehav. Rev 37, 681–696. doi:10.1016/j.neubiorev.2013.02.002. [PubMed: 23415703]

Shankman SA, Funkhouser CJ, Klein DN, Davila J, Lerner D, Hee D, 2017. Reliability and validity of severity dimensions of psychopathology assessed using the structured clinical interview for DSM-5 (SCID). Int. J. Methods Psychiatr. Res 27. doi:10.1002/mpr.1590.

Shaw DS, Gilliom M, Ingoldsby EM, Nagin DS, 2003. Trajectories leading to school-age conduct problems. Dev. Psychol 39, 189–200. doi:10.1037//0012-1649.39.2.189. [PubMed: 12661881]

Shaw DS, Hyde LW, Brennan LM, 2012. Early predictors of boys' antisocial trajectories. Dev. Psychopathol 24, 871–888. doi:10.1017/S0954579412000429. [PubMed: 22781860]

Spreckelmeyer KN, Krach S, Kohls G, Rademacher L, Irmak A, Konrad K, Kircher T, Gründer G, 2009. Anticipation of monetary and social reward differently activates mesolimbic brain structures in men and women. Soc. Cogn. Affect. Neurosci 4, 158–165. doi:10.1093/scan/nsn051. [PubMed: 19174537]

Stoffel MA, Nakagawa S, Schielzeth H, 2017. rptR: repeatability estimation and variance decomposition by generalized linear mixed-effects models. Methods Ecol. Evol 8, 1639–1644. doi:10.1111/2041-210X.12797.

Streiner DL, Norman GR, 2008. Health Measurement Scales. Oxford University Press doi:10.1093/acprof:oso/9780199231881.001.0001.

Sutherland MT, Stein EA, 2018. Functional neurocircuits and neuroimaging biomarkers of tobacco use disorder. Trends Mol. Med., Biomark. Substance Abuse 24, 129–143. doi:10.1016/j.molmed.2017.12.002.

Taylor BK, Frenzel MR, Eastman JA, Wiesman AI, Wang Y-P, Calhoun VD, Stephen JM, Wilson TW, 2020. Reliability of the NIH toolbox cognitive battery in children and adolescents: a 3-year longitudinal examination. Psychol. Med 1–10. doi:10.1017/S0033291720003487.

Telzer EH, McCormick EM, Peters S, Cosme D, Pfeifer JH, van Duijvenvoorde ACK, 2018. Methodological considerations for developmental longitudinal fMRI research. Dev. Cogn. Neurosci 33, 149–160. doi:10.1016/j.dcn.2018.02.004. [PubMed: 29456104]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Thomas DR, Zumbo BD, 2012. Difference scores from the point of view of reliability and repeated-measures ANOVA: in defense of difference scores for data analysis. Educ. Psychol. Meas 72, 37–43. doi:10.1177/0013164411409929.

van den Bulk BG, Koolschijn PCMP, Meens PHF, van Lang NDJ, van der Wee NJA, Rombouts SARB, Vermeiren RRJM, Crone EA, 2013. How stable is activation in the amygdala and prefrontal cortex in adolescence? A study of emotional face processing across three measurements. Dev. Cogn. Neurosci., Special Issue: Neural Plast., Behav. Cogn. Train.: Dev. Neurosci. Perspect 4, 65–76. doi:10.1016/j.dcn.2012.09.005 .

Van Essen DC, Glasser MF, Dierker DL, Harwell J, Coalson T, 2012. Parcellations and hemispheric asymmetries of human cerebral cortex analyzed on surface-based atlases. Cereb. Cortex 22, 2241–2262. doi:10.1093/cercor/bhr291. [PubMed: 22047963]

Vetter NC, Steding J, Jurk S, Ripke S, Mennigen E, Smolka MN, 2017. Reliability in adolescent fMRI within two years – a comparison of three tasks. Sci. Rep 7, 1–11. doi:10.1038/s41598-017-02334-7 . [PubMed: 28127051]

Wiggins GC, Polimeni JR, Potthast A, Schmitt M, Alagappan V, Wald LL, 2009. 96-channel receive-only head coil for 3 Tesla: design optimization and evaluation. Magn. Reson. Med. Off. J. Soc. Magn. Reson. Med. Soc. Magn. Reson. Med 62, 754–762. doi:10.1002/mrm.22028.

Wu CC, Samanez-Larkin GR, Katovich K, Knutson B, 2014. Affective traits link to reliable neural markers of incentive anticipation. Neuroimage 84, 279–289. doi:10.1016/j.neuroimage.2013.08.055. [PubMed: 24001457]

Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD, 2011. Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods 8, 665–670. doi:10.1038/nmeth.1635. [PubMed: 21706013]

Yeo BTT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, Roffman JL, Smoller JW, Zöllei L, Polimeni JR, Fischl B, Liu H, Buckner RL, 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol 106, 1125–1165. doi:10.1152/jn.00338.2011. [PubMed: 21653723]

Zink CF, Pagnoni G, Martin-Skurski ME, Chappelow JC, Berns GS, 2004. Human striatal responses to monetary reward depend on saliency. Neuron 42, 509–517. [PubMed: 15134646]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

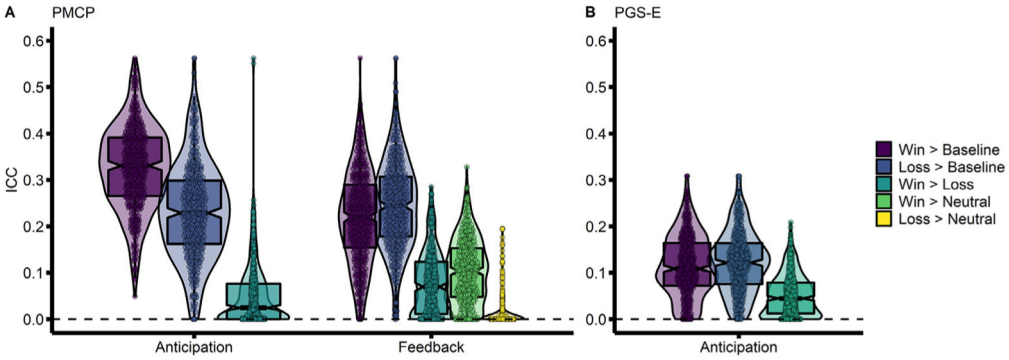

Fig. 1. Win > Baseline contrast has a higher stability than Win > Loss contrast

Stability of activation across the whole brain, using the Schaefer-400 cortical parcellation. All five contrasts for examined task phases (anticipation and feedback) are shown. (A) The Pitt Mother & Child Project (PMCP; $N = 139$) sample. (B) The Pittsburgh Girls Study – Emotions substudy (PGS-E; $N = 145$) sample. Associated statistics are reported in the main text, as well as in Supplemental Table 1. .

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

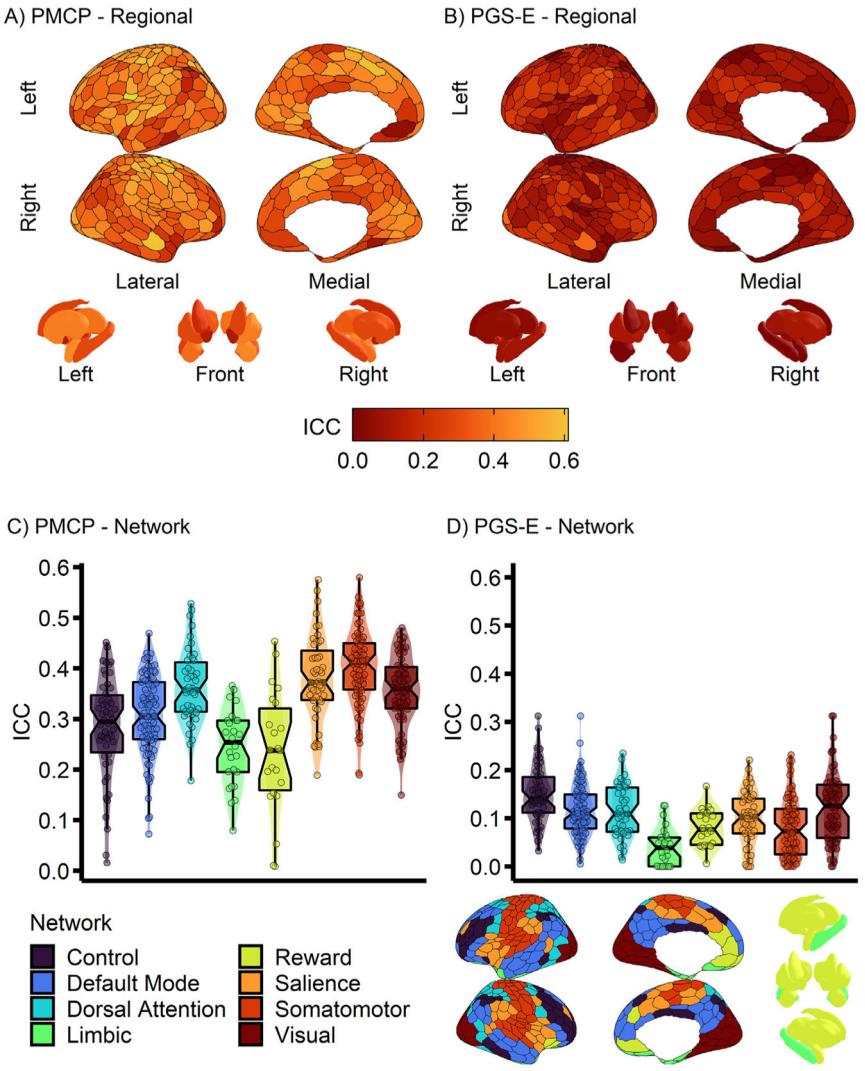

Fig. 2. Stability is lower in the limbic and reward networks

Stability of activation during the anticipation of monetary gains (Win>Baseline contrast), using the Schaefer-400 cortical parcellation and Harvard-Oxford subcortical atlas, across all networks examined. (A&B) ROI reliability visualized on a surface projection. (A&C) the Pitt Mother & Child Project (PMCP; $N = 139$) study, and (B&D) the Pittsburgh Girls Study–Emotions substudy (PGS-E; $N = 145$). (C&D) Associated statistics are reported in Supplemental Tables 2 and 3. .

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

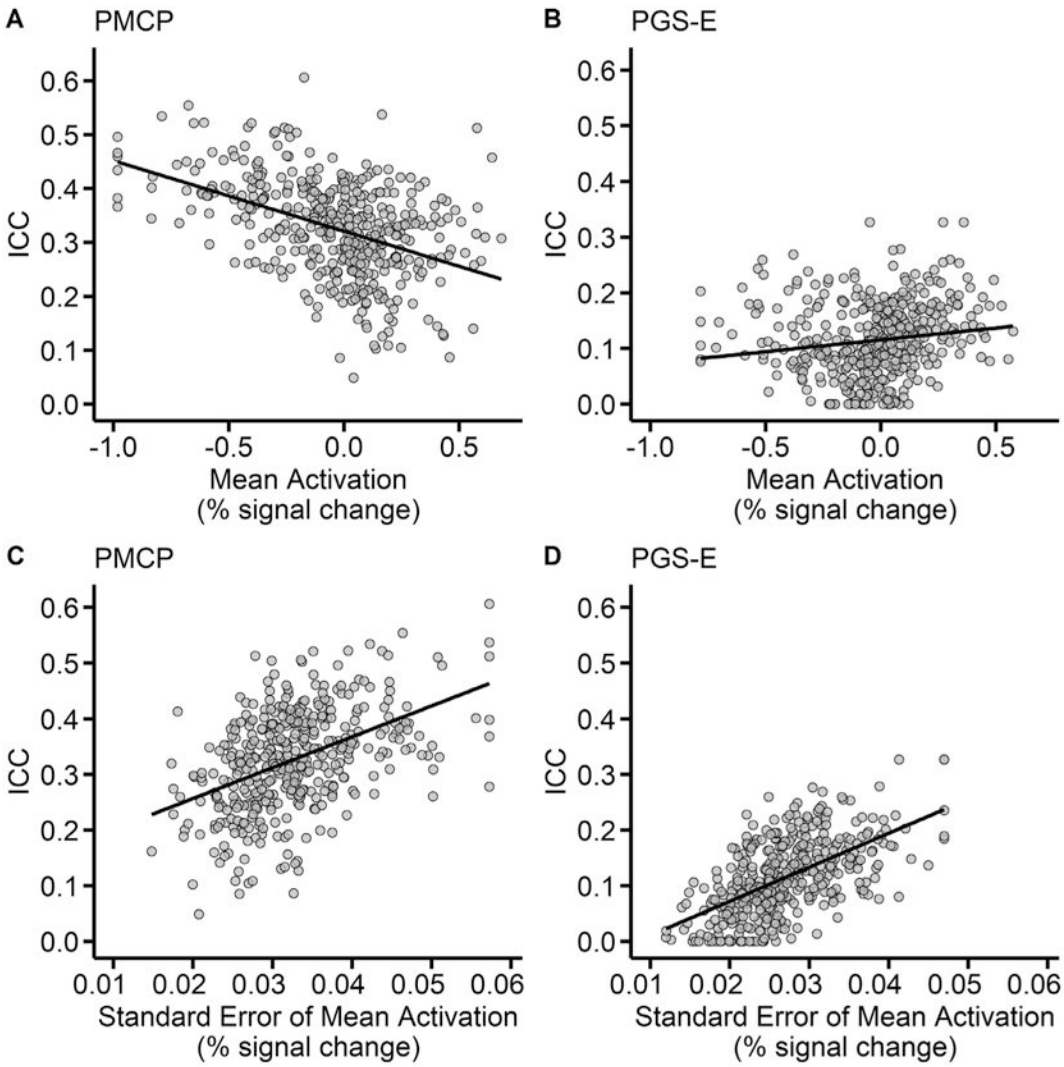

Fig. 3. Regions with greater between-subject variation are more stable

Correlations between the stability of activation during the anticipation of monetary gains (Win> Baseline contrast), and region mean activation (A&B) and variability of activation (C&D). (A&B) the Pitt Mother & Child Project (PMCP; $N = 139$) study, and (C&D) the Pittsburgh Girls Study– Emotions substudy (PGS-E; $N = 145$). .

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Sample demographics.

PMCP	***N* = 139 (*n* = 278 scans; 2 scans per participant)**						
Age by wave	**Wave 1** 20.06 (0.21)	**Wave 2** 22.04 (0.13)					
Race	**White American** *n* = 72 (51.8%)	**African American** *n* = 55 (39.57%)	**Mixed** *n* = 8 (5.76%)	**Other** *n* = 4 (0.72%)			
Time between scans (years)	**Mean (SD)** 1.98 (0.23)	**Median (IQR)** 2 (1.95–2.06)	**Min** 0.77	**Max** 2.7			
PGS-E	***N* = 145 (*n* = 439 2–4 scans per participant)**						
Number of scans	**2 scans** *n* = 45 (31.1%)	**3 scans** *n* = 51 (35.17%)	**4 scans** *n* = 49 (33.79%)				
Age by wave	**Wave 1** 16.85 (0.51)	**Wave 2** 18.01 (0.5)	**Wave 3** 19.18 (0.47)	**Wave 4** 20.28 (0.46)			
Age at scan	**15 yrs. old** *n* = 7 (1.6%)	**16 yrs. old** *n* = 61 (13.9%)	**17 yrs. old** *n* = 101 (23%)	**18 yrs. old** *n* = 101 (23%)	**19 yrs. old** *n* = 101 (23%)	**20 yrs. old** *n* = 61 (13.9%)	**21 yrs. old** *n* = 7 (1.6%)
Race	**White American** *n* = 35 (24%)	**African American** *n* = 101 (69.66%)	**Mixed** *n* = 9 (4.83%)				
Time between scans (years)	**Mean (SD)** 1.27 (0.54)	**Median (IQR)** 1.157 (0.99–1.33)	**Min** 0.44	**Max** 4.18			

Descriptive information of the sample demographics of the Pitt Mother & Child Project (PMCP) study and the Pittsburgh Girls Study – Emotions substudy (PGS-E). SD= standard deviation. IQR = inter quartile range (Q1 – Q3).