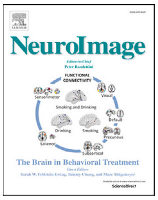
Contents lists available at ScienceDirect

NeuroImage

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



Test-retest reliability of fMRI-measured brain activity during decision making under risk

Ozlem Korucuoglu[a,*], Michael P. Harms[a], Serguei V. Astafiev[a], James T. Kennedy[a], Semyon Golosheykin[a], Deanna M. Barch[a,b], Andrey P. Anokhin[a]

[a] Department of Psychiatry, Washington University School of Medicine, 660 S. Euclid, St. Louis, MO, 63110, USA
[b] Department of Psychological & Brain Sciences, Washington University, 1 Brookings Drive, St. Louis, MO, 63130, USA

ARTICLE INFO

Keywords:
BART
Cohen's d
Familiality
fMRI
ICC
Motion

ABSTRACT

Neural correlates of decision making under risk are being increasingly utilized as biomarkers of risk for substance abuse and other psychiatric disorders, treatment outcomes, and brain development. This research relies on the basic assumption that fMRI measures of decision making represent stable, trait-like individual differences. However, reliability needs to be established for each individual construct. Here we assessed long-term test-retest reliability (TRR) of regional brain activations related to decision making under risk using the Balloon Analogue Risk Taking task (BART) and identified regions with good TRRs and familial influences, an important prerequisite for the use of fMRI measures in genetic studies. A secondary goal was to examine the factors potentially affecting fMRI TRRs in one particular risk task, including the magnitude of neural activation, data analytical approaches, different methods of defining boundaries of a region, and participant motion. For the average BOLD response, reliabilities ranged across brain regions from poor to good (ICCs of 0 to 0.8, with a mean ICC of 0.17) and highest reliabilities were observed for parietal, occipital, and temporal regions. Among the regions that were of *a priori* theoretical importance due to their reported associations with decision making, the activation of left anterior insula and right caudate during the decision period showed the highest reliabilities (ICCs of 0.54 and 0.63, respectively). Among the regions with highest reliabilities, the right fusiform, right rostral anterior cingulate and left superior parietal regions also showed high familiality as indicated by intrapair monozygotic twin correlations (ranging from 0.66 to 0.69). Overall, regions identified by modeling the average BOLD response to a specific event type (rather than its modulation by a parametric regressor), regions including significantly activated vertices (compared to a whole parcel), and regions with greater magnitude of task-related activations showed greater reliabilities. Participant motion had a moderate negative effect on TRR. Regions activated during decision period rather than outcome period of risky decisions showed the greatest TRR and familiality. Regions with reliable activations can be utilized as neural markers of individual differences or endophenotypes in future clinical neuroscience and genetic studies of risk-taking.

1. Introduction

Individuals differ in their preferences to engage in behaviors that involve a certain amount of possible risks and rewards. Neuroimaging research has attempted to unravel the neural basis of these risk attitudes in an effort to understand human behavior. Neural correlates of decision making under risk have been utilized as a brain-based biomarker of treatment outcomes (Chung et al., 2009; Macoveanu et al., 2014), investigated as a likely heritable trait (Rao et al., 2018), and examined across development (Qu et al., 2015b) to aid our understanding of both typical and aberrant decision processes. Moreover, recent research has started to focus on parametrically modulated neural correlates of decision processes across development in relation to the probability and magnitude of the choices (Insel and Somerville, 2018; Korucuoglu et al., 2019).

A key (and often implicit) assumption of the above lines of research is that neural correlates of risk-taking represent reliable, trait-like measures. However, without confidence in the stability of individual

Abbreviations: BART, Balloon Analogue Risk Task; fMRI, functional Magnetic Resonance Imaging; ICC, Intraclass correlation.

[*] Corresponding author. Department of Psychiatry, Washington University School of Medicine, 660 S. Euclid, Campus Box 8134, St. Louis, MO, 63110, USA.
E-mail addresses: korucuogluo@wustl.edu, ozlemkorucuoglu@gmail.com (O. Korucuoglu).

https://doi.org/10.1016/j.neuroimage.2020.116759
Received 23 September 2019; Received in revised form 1 February 2020; Accepted 15 March 2020
Available online 20 March 2020

differences in these measures over time (also known as test-retest reliability, TRR), the differences across measurements cannot be attributable to the impact of treatment or developmental changes. Using paradigms other than risk taking, previous TRR studies of task-fMRI showed modest reliabilities (Bennett and Miller, 2013; Gorgolewski et al., 2013a, 2013b). A recent meta-analysis of 11 fMRI tasks concluded that commonly used task-fMRI measures have poor reliability and therefore may be unsuitable for individual differences research (Elliott et al., 2019). More importantly, the range of these reliabilities varies greatly across different constructs, making generalization very difficult (Elliott et al., 2019; Frohner et al., 2019). Therefore, reliabilities of neural activations need to be estimated for each specific construct and task. Lastly, investigating systematic changes in the brain responses to increasing intensity of a stimulus (i.e., degree of risk/reward) requires a design with parametric modulation. While parametrical fMRI data analytical approaches could provide valuable information for updating current decision making models (especially in adolescents), the reliability of parametrically modulated neural responses has yet to be established. Although decision making under risk is a highly studied construct in neuroscience, we are not aware of published studies that investigated the TRRs of the neural activations underlying decision making under risk.

Another important issue is to determine factors that affect TRRs. Given that much fMRI research has historically focused on group-level activations, some studies tested the reliability of group activation maps (Caceres et al., 2009; Raemaekers et al., 2007). Although Caceres and colleagues' study (2009) demonstrated a higher probability of greater TRRs within highly activated regions in the first session, regions with low activations and high reliabilities were also identified. However, small sample sizes in the aforementioned studies (10 and 12 subjects), might bias the chance of identifying genuine differences across regions or as a function of other factors in reliability estimates. To determine the extent to which the group-level magnitude of regional activation predicts reliability of individual differences, in the present study we examined the relationship between the magnitude of regional activation in the first scanning session and the TRR of that activation across brain regions.

Furthermore, previous studies of task-fMRI, especially the ones that studied brain based biomarkers of disease, focused on a selected number of regions of interest (ROIs). Physical boundaries of these ROIs are defined differently across different studies (utilizing readily available parcellation schemes vs. defining the boundaries using the voxels (or vertices) above a certain threshold, i.e., 'significant voxels'), each having their advantages and disadvantages. The use of a parcellation scheme can provide easy comparison of results across studies and scans for the same voxels, which is an advantage of favoring parcellation schemes over activation based ROIs, though only if the inclusion of below-threshold voxels to the ROIs do not negatively affect the reliabilities. However this assumption needs to be investigated empirically. Another consideration in favor of investigating reliabilities for regions in a parcellation scheme in addition to regions identified through significant group-level activations is that a brain region may be activated only in a subset of subjects, while other subjects may show no activation or even deactivation in the same region, leading to modest activation at the group level that may not reach statistical significance. Nevertheless, individual differences in such regions may be reliable. Thus, in an approach focusing only on significant group-level activations, such regions showing reliable and potentially meaningful individual differences in activation may be missed.

Lastly, signal variability and artifacts due to motion is another factor that impacts the data quality and reliability. Motion itself varies across different samples (children, adolescents, adults; patients vs controls), and task designs (task type, scanning duration) (Engelhardt et al., 2017; Zeng et al., 2014).

In the current study our main goal was to assess long-term (i.e., over 6 months) test-retest reliability of regional brain activations related to decision making under risk with the use of the Balloon Analogue Risk Taking Task (BART) in a community-representative adult sample composed of monozygotic twins (MZ). Based on a systematic review of the fMRI literature utilizing BART (See Supplementary Materials Table S2), we focused on contrasts that may play a primary (decision making) or secondary (outcome evaluations) role in risky decision making, given that different studies may prefer to implement different contrasts dependent on study specific aims. Note that this selection is not exhaustive given that individual studies utilized other contrasts to achieve their distinct goals, e.g. to investigate processes related to loss aversion (Fukunaga et al., 2012) or bias in decision making due to prior outcomes (Kohno et al., 2015). We provide TRRs separately on the whole-brain parcel level, for significant activations, and for a set of selected parcels. We expected poor to modest reliabilities based on earlier reliability studies utilizing different constructs (Bennett and Miller, 2010; Elliott et al., 2019).The inclusion of MZ twins allowed us to obtain preliminary evidence for familial transmission.[1] Evidence for test-retest reliability is an important prerequisite for genetic studies because only trait-like individual differences can be heritable, and test-retest reliability can be viewed as the upper bound for heritability (Mccrae et al., 2011). To that end, we examined whether test-retest reliability of regional activations predicted the size of familial influences on fMRI measures on the whole brain parcel level and aimed to identify regions with both high reliability and familiality that can be targeted as candidate endophenotypes in genetic studies. Our second goal was to examine the factors potentially affecting TRRs including the strength (magnitude) of regional activation, data analytical approaches (categorical versus parametric), different methods of defining boundaries of a region (by using a parcellation scheme versus using significant vertices within a parcel), and motion.

2. Methods

2.1. Participants

Fifty-six young adults (32 females, age range: 21–24 years, mean = 23.27, SD = 0.86) participated in the study. Participants were monozygotic (MZ) twins ascertained through the Missouri Family Registry maintained at the Department of Psychiatry at Washington University School of Medicine (WUSM) as part of a larger study - Genetics, Neurocognition, and Adolescent Substance Abuse (GNASA). All 56 participants in the present study completed the first MRI scanning session (Time1), and 44 of them (26 females,[2] age range: 21–24 years, mean = 23.31, SD = 0.89) completed a second session approximately 6 months later (Time2, mean interval 7.9 months, ranging from 5.7 to 12.0 months).[3] Exclusion criteria included (1) standard MRI contraindications such as non-removable metal in the body, dental braces, excessive weight, claustrophobia, current pregnancy, or difficulty lying supine; (2) intellectual or physical impairments or uncorrectable sensory impairment precluding participation in the laboratory session, (3) known diagnoses of schizophrenia, autism, bipolar disorder, or epilepsy[4] since these disorders are known to be associated with specific cognitive impairments that may interfere with the administration of experimental tests; (4) inability to understand English; and (5) history of head trauma with loss of consciousness for more than 5 min. Before inclusion to the study,

[1] We use the term "familiality" (Kendler and Neale, 2009) because MZ twins, by themselves, are not sufficient to distinguish between heritable (genetic) and environmental effects.

[2] Reliability estimates of the parcels covering the entire brain (*unthresholded parcels* analysis) for groups composed of all males and all females yielded very similar values (see *Supplementary Materials* Figure S1).

[3] We investigated a possible influence of the time interval between the two scanning sessions on the between session variability in beta estimates for the parcels covering the entire brain (*unthresholded parcels* analysis) and did not find any relationship (see *Supplementary Materials* Figure S2).

[4] History of depression was not assessed, although depression has been associated with moderate cognitive impairment as well (Rock et al., 2014).

participants were screened for these exclusion criteria via self-report. Upon arrival to the lab, they also completed a urine drug test [for Methamphetamine, Opiates, PCP, Benzodiazepines, Methadone, Barbiturates, Amphetamines, Cocaine, TetraHydroCannabinol (THC)] and an alcohol breathalyzer test. One participant's session was rescheduled because of a positive drug test for THC. The Human Research Protection Office at the Washington University School of Medicine approved the study. A written informed consent was obtained from all participants. Participants were compensated for participation in the study.

2.2. In-scanner balloon analogue risk task (BART) description

We used a scanner version of BART modified by Rao et al. (2008) (also see Fig. 1). Before the actual scanning, participants were placed in a mock scanner for accommodation to the scanner environment, where they received instructions and performed a practice version of the in-scanner tasks (see *Supplementary Materials* for further details). In the BART paradigm, participants were given the chance to earn money by sequentially inflating a balloon without popping it. A maximum of 12 inflations were possible for each balloon with the probability of explosion and possible earnings increasing monotonically (see *Supplementary Materials*, Table S1 for probability of explosions and possible earnings by number of inflations). All balloons had the same sequence of explosion probabilities. With each inflation participants could earn additional money or at any time they could stop inflating the balloon and cash-out the amount accumulated for the current balloon into a virtual bank. However, the balloons could explode unpredictably at varying degrees of inflation, in which case the accumulated gain for the current balloon would be lost (but the amount that had previously been cashed-out into the bank was unaffected). Thus, this task entailed an approach-avoidance conflict, such that each subsequent inflation increased the total amount of possible gain while, at the same time, the risk of losing the accumulated gain of that trial increased as well.

The total task duration was set to 10 min (acquired over a single run), during which participants completed as many trials as possible (variable called *Balloons completed*). The task started with a fixation period of 30 s. A trial started with a balloon and a green rectangular cue, during which subjects had unlimited time to respond (a button press with index finger to pump the balloon or with the middle finger to cash-out). Following the response, the balloon remained on the screen for 0, 2, 4, or 6 s during which the balloon size did not change. The duration of the delay following the pump was randomly decided and each delay interval was given an exponentially decreasing weight (weights were 30, 12, 5, and 2, respectively, for the delay intervals of 0, 2, 4, and 6 s). The participant's response led to 3 possible outcomes: (1) if the participant cashed-out, the text "You Win" was presented for 1 s; (2) if the participant pumped the balloon and the balloon exploded, an exploded balloon was presented for 0.5 s, followed by the text "You Lose" for 1 s; and (3) if the balloon inflated successfully, the color of the rectangular cue switched to red for an equiprobable 1.5, 2, or 2.5 s). During the red cue period, subjects were instructed not to give any response. After explosions or cash-outs, but before the next balloon appeared on the screen, a blank screen was presented for an equiprobable 2, 3, or 4 s (the inter-stimulus interval; ISI). The value of the current pump was displayed on the balloon and the total amount of winnings across task was displayed under the rectangular cue at all times when the balloon was visible. Participants were paid their earnings at the end of the task as an extra bonus (average earning amount for Time1 = $14.78 across 56 participants and for Time2 = $15.46 across 44 participants), in addition to the compensation for study participation.

During each scanning session, participants performed six cognitive tasks in a predetermined order, with the BART task presented as the 3rd one. The first and the second tasks lasted 12 min each.

2.3. fMRI data acquisition

Echo-planar imaging (EPI) of the whole brain was acquired with a 32 channel head coil on a 3T Siemens MAGNETOM Prisma scanner in the WUSM Neuroimaging Labs, using Human Connectome Project (HCP) style acquisitions. The specific sequence implementations and scanning parameters were the same as those used for the Adolescent Brain Cognitive Development (ABCD) Study (Casey et al., 2018). Structural scans included a sagittal magnetization prepared gradient-echo (MP-RAGE) T1-weighted image (repetition time [TR] = 2500 msec; echo time [TE] = 2.88 msec; flip angle = 8^0; voxel size = 1.0 x 1.0 × 1.0 mm) and a sagittal T2-weighted image (T2-SPACE, TR = 3200 msec; TE = 565 msec; voxel-size = 1.0 x 1.0 × 1.0 mm). Both the T1w and T2w scans utilized embedded volumetric navigators that detected and compensated for head movement in real-time, with an allowance for reacquisition of the lines (TRs) in *k*-space that are heavily corrupted by motion (up to 24 TRs for the MP-RAGE, and 18 TRs for the T2-SPACE scan). The combination of real-time motion correction and *k*-space reacquisition improves the quality of the structural scans and reduces the need for rescans, especially for age groups with a higher incidence of head movement (Tisdall et al., 2012). BOLD contrast for the task was measured with a gradient-echo EPI sequence (TR = 800 msec; TE = 30 msec; 775 frames; 60 contiguous 2.4 mm transverse slices; 2.4 × 2.4 mm in plane resolution, multi-band factor 6, posterior-to-anterior phase encoding). Two brief spin-echo EPI scans with opposite phase-encoding directions (anterior-posterior and posterior-anterior) were acquired immediately before the BOLD scan for the purpose of correcting susceptibility distortion.

2.4. fMRI data processing

The HCP data analysis pipelines (https://github.com/Washington-University/HCPpipelines, v.3.19.0) were used for the analysis of fMRI images (Glasser et al., 2013). The following pipelines were used: three structural preprocessing pipelines (*PreFreeSurfer*, *FreeSurfer*, and *PostFreeSurfer*), and two functional pipelines (*fMRIVolume* and *fMRISurface*). The main purpose of the *PreFreeSurfer* pipeline is to generate an

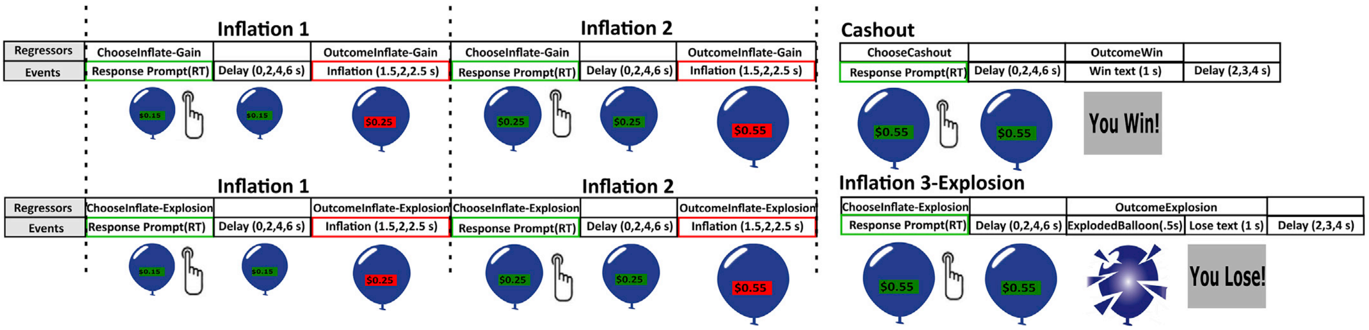


Fig. 1. Schematic representation of the Balloon Analogue Risk Task (BART). Cashed-out (upper panel) and Exploded balloon trials (lower panel) are depicted in the figure. Participants were instructed to give their response during the green rectangular cue period and participant's response triggers the onset of delay period.

undistorted “native” structural volume space for each subject, align the T1w and T2w images, perform B1 (receive-coil bias field) correction, and register the subject’s native structural volume space to MNI space. The *FreeSurfer* pipeline used FreeSurfer version 5.3.0-HCP. The main purpose of this pipeline is to construct white and pial cortical surfaces, compute FreeSurfer’s standard folding-based surface registration, and segment the subcortical structures. Finally, the *PostFreeSurfer* pipeline produces all of the NIFTI volume and GIFTI surface files necessary for viewing the data in Connectome Workbench, creates myelin maps, and applies the surface registration (including down-sampling to a lower resolution, common mesh). Surface registration across subjects used FreeSurfer’s standard folding-based registration – ‘MSMSulc’ registration (a more gentle folding-based alignment with less distortion (Robinson et al., 2018)) was not used because the necessary ‘msm’ binary was not publicly available at the time we started processing. Following the structural pipelines, all data underwent careful quality control (see *Supplementary Materials*). The *fMRIVolume* preprocessing pipeline includes correction for gradient nonlinearities, volume realignment to compensate for subject motion, EPI distortion correction, bias field reduction, brain-boundary-based registration of EPI to structural T1-weighted scan, non-linear (FNIRT) registration into MNI152 space, grand-mean intensity normalization and masking the data with the final brain mask. The *fMRISurface* pipeline transforms the time series from the volume into a CIFTI (Connectivity InFormatics Technology Initiative) grayordinate standard space (a space containing cortical gray matter surface vertices, and subcortical gray matter volume voxels, but excluding white matter and CSF; allowing combined cortical surface and subcortical volume analysis (Glasser et al., 2013)). Surface-based registration for the cortical data improves the alignment of task-evoked data across subjects (Coalson et al., 2018). The HCP *TaskfMRIAnalysis* pipeline, which uses FEAT tool (FMRIB’s Expert Analysis Tool) from FSL v6.0 (Jenkinson et al., 2012), was used to analyze the cortical and subcortical grayordinate data for task modeling. The first eight frames were discarded from further analysis to allow for equilibrium of the longitudinal magnetization.

For task modeling, we used two distinct approaches, which were based on previous studies utilizing BART in the scanner, so that we could report reliabilities for well-studied contrasts: categorical modeling of BOLD responses to different event types (*categorical design*) and parametric modeling in which the probability of explosion was used as parametric modulator (*parametric design*).[5] In the results section we report TRR estimates for both the main and modulator regressors, however due to lower reliabilities of brain activations obtained using the modulator regressors, the main focus of the discussion is on the analysis using the categorical modeling of BOLD responses. Fig. 1 demonstrates the sequence of events and EVs in the task for cashed-out and exploded balloons. The categorical model included 3 choice related and 4 outcome related regressors. Choice related regressors included ‘ChooseInflate-Gain’ and ‘ChooseInflate-Explosion’ regressors preceding pumps – one for balloons that were subsequently cashed-out (gain) and one for balloons that were subsequently exploded (explosion) – and a ‘ChooseCashout’ regressor. Outcome related regressors included ‘OutcomeExplosion’ and ‘OutcomeWin’ regressors, plus ‘OutcomeInflate-Gain’ and ‘OutcomeInflate-Explosion’ regressors for successful pumps, for balloons that were subsequently cashed-out vs. balloons that subsequently exploded, respectively. The ‘ChooseInflate-Explosion’ and ‘OutcomeInflate-Explosion’ events were included in the model as ‘conditions of no interest’. The reason that the ‘ChooseInflate-Explosion’ and ‘ChooseInflate-Gain’ events were modeled with separate regressors is that using a single regressor to model all pumps preceding explosions would have resulted in the inclusion of trials in which participants were forced to stop pumping because of the explosion itself.

Choice related regressors were modeled with a duration (prior to convolution with the hemodynamic response function) equal to the interval from the onset of the green rectangular cue until the response. The ‘ChooseInflate-Gain’ regressor preceding cash-outs included all pumps except the cash-out button press. Similarly, the ‘ChooseInflate-Explosion’ regressor included all pumps before the explosion; this included the last inflated balloon presentation before explosion. ‘OutcomeInflate-Gain’ and ‘OutcomeInflate-Explosion’ regressors were modeled with a duration equal to the red rectangle cue presentation. The ‘OutcomeExplosion’ regressor included the duration of the presentation of exploded balloon plus the presentation of the ‘You Lose’ feedback (i.e., 1.5 s total). The duration of the ‘OutcomeWin’ regressor was always 1 s (the duration of the ‘You Win’ feedback).

The categorical model included 4 contrasts, each defined from a single regressor, thus each representing a comparison to the baseline (fixation periods at the beginning and at the end of the task (each 30 s), as well as the delay periods following inflations, cashouts, explosions and win outcome). These were: (1) *ChooseInflate* (preceding cash-outs; i.e., the ChooseInflate-Gain regressor); (2) *ChooseCashout*; (3) *OutcomeInflate* (the presentation of inflated balloon, preceding cash-outs; i.e., OutcomeInflate-Gain regressor)); (4) *OutcomeExplode* (i.e., OutcomeExplosion regressor).

In the parametric model, the probabilities of explosions [P(explode)] were included as a parametric modulator with each event type (EV) regressor. The parametric model included all the same regressors and contrasts as the categorical model, except the ‘OutcomeWin’ regressor was included as non-parametric regressor, because the probability of explosion was no longer applicable at this point.

Group level grayordinate-wise statistical maps were created, for Time1 and Time2 separately, by using permutation statistics as implemented in the PALM toolbox, version alpha101 (Permutation Analysis of Linear Models, http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/PALM Winkler et al., 2014), using just the participants that completed scanning at both time points ($n = 44$). Multi-level exchangeability blocks (Winkler et al., 2015), which limit the permutations within block level (i.e., between two MZ siblings), were used to account for the shared variance between twins. False discovery rate (FDR) corrected activation maps for all contrasts separately for Time1 and Time2 ($n = 44$) are provided in the *Supplementary Materials* (see Figure S3 and Figure S4, for significant activations in the categorical and parametric designs, respectively). We did not correct for the multiplicity of contrasts.

2.5. Parcellation/segmentation of the fMRI data

Test-retest reliabilities were calculated on the mean BOLD response magnitudes (“beta weights”, i.e., contrast estimates computed by *TaskfMRIAnalysis* pipeline) extracted from the following three sets of ROIs: (1) the whole-brain grayordinates divided into 360 cortical parcels and 19 subcortical segmentations without regard to activation level (referred to as *unthresholded parcels*); (2) activations significant at the group level divided into anatomical parcels (872 and 463 parcels in the categorical and parametric designs, respectively) (referred to as *thresholded parcels*); (3) a subset of the *thresholded parcels* (106 and 24 parcels in the categorical and parametric designs, respectively), representing the regions that are most prominent in the decision making literature (referred to as the *thresholded subset*). *Unthresholded parcels* were used to study the relationship between familiality and TRRs and to investigate potential factors that affect the TRRs (i.e., magnitude of activation at Time1, categorical versus parametric fMRI analytical approaches, motion, tSNR). The *thresholded parcels* and *thresholded subset* were used to investigate reliabilities in task related active regions and to identify

[5] Explosion probabilities were used as the parametric modulator in the current study, consistent with previous BART studies using the same task (Fukunaga et al., 2012; Hulvershorn et al., 2015; Rao et al., 2008). Using reward value as the modulator would be another option, however given that the correlation (Pearson r) between P[explosion] and reward value was 0.99, both modeling approaches would effectively yield the same results. Note that pump number has also been used as the parametric modulator in some of the previous research using a different scanner version of BART (Kohno et al., 2015; Qu et al., 2015a).

potential endophenotypes that can be targeted in future genetic studies. Moreover, to determine whether different approaches to defining ROIs affect test-retest reliabilities, we compared TRRs of the *unthresholded parcels* with the TRRs of the *thresholded parcels*. This allowed us to meaningfully compare results from whole parcels (i.e., without regard to activation) to parcels that included only group level significant grayordinates. This comparison was tested for all parcels that contained significant activations.

2.5.1. Parcellation/segmentation of all grayordinates (unthresholded parcels)

Grayordinate-wise beta weights were divided into 360 parcels using the Human Connectome Project Multi-Modal Parcellation, version 1.0 (MMP1.0, Glasser et al., 2016) and the Freesurfer-derived 19 structure subcortical segmentation (Fischl et al., 2002) embedded into the CIFTI output by the HCP Pipelines.

2.5.2. Parcellation/segmentation of significant group activation maps (thresholded parcels)

Based on the Time1 maps (n = 44), the whole-brain grayordinate-wise FDR corrected maps were used to define significant clusters for each contrast as groups of spatially contiguous grayordinates exceeding 80mm^2/120 mm^3 (surface/volume for cortical and subcortical regions,[6] respectively). The FDR corrected significant clusters were further divided into anatomical parcels using the parcellation/segmentation scheme described above. Throughout the text, the term "*thresholded parcels*" is used to refer to the conjunction between a parcel/segment and the FDR corrected significant clusters. Thus, all thresholded activations respected the parcellation boundaries (and were entirely inside one of the parcels), but only contained above threshold vertices/voxels.

2.5.3. Selection of the thresholded activation maps (thresholded subset)

The purpose of selecting a subset of regions was to report reliabilities for regions that are of particular theoretical importance due to their reported associations with decision making processes. In this selection process, we applied a stepwise procedure: starting with the *thresholded parcels*, we first identified all regions that had a moderate-to-high effect size (for a complete description of selection protocol, see below); then we further selected the subset of those regions that were reported in two meta-analyses of risk taking/decision making [regions listed in Table 2 and 4 of Krain et al. (2006) study; regions listed in Table 2, 3, and 4 of Silverman et al. (2015) study] or in previous BART fMRI studies (for a complete list of these regions, see Supplementary Table S2).

Specifically, among the *thresholded parcels* at Time1, regions were selected if they (a) had a Cohen's d value > 0.35 (small/medium effect size) for cortical and >0.2 (small) for subcortical regions for the categorical design and Cohen's d > 0.2 for the cortical and subcortical regions identified by the parametric design (according to Cohen's effect size classification, d = 0.2 is a "small" effect, and d = 0.35 is midway between the small and the "medium" effect of d = 0.5)[7]; (b) had significant activation present in >50% of the parcel/segment[8]; (c) had been reported in previously published risk taking/decision making meta-analyses or fMRI studies of the BART. Any regions in the primary visual cortex were excluded as non-task specific regions of activation consistently observed in most visual tasks. Notably, none of the subcortical segments passed criteria (b) in either design (categorical or parametric), and thus the only subcortical data is from the *unthresholded* and *thresholded parcels* analyses.

Table 1
Summary statistics for the behavioral outcome variables of the in-scanner BART task.

Variables	Paired Samples T-Test Results				TRR	Familiality	
	N[a]	Time1 (m, SD)	Time2 (m, SD)	*p*	(ICC)	N[a]	ICC
Balloons completed	42	21.86 (2.47)	21.81 (2.45)	.75	.56*F	27	.12
N pumps per balloon	43	5.55 (.74)	5.58 (.71)	.72	.62*G	27	.06
RT pumps (ms)	44	794.72 (286.29)	767.46 (282.66)	.41	.71*G	27	.23
% Explosion Rate	44	40.18 (13.41)	38.88 (16.40)	.57	.50*F	27	.15
N Cash-outs	43	13.30 (4.19)	13.49 (4.61)	.75	.62*G	27	.17
N pumps preceding cash-outs	44	5.58 (.88)	5.60 (.89)	.84	.67*G	27	.07
RT cash-outs (ms)	44	780.73 (400.16)	681.65 (323.75)	**.02**	.73*G	26	.47*
RT pumps preceding cash-outs	44	798.65 (285.91)	776.06 (284.17)	.53	.65*G	27	.25
N Explosions	44	8.68 (2.68)	8.39 (3.26)	.55	.40*F	27	.08
N pumps preceding explosions	43	5.51 (.85)	5.48 (.72)	.81	.42*F	27	.10
RT explosions (ms)	42	760.20 (323.00)	712.63 (293.63)	.32	.52*F	26	.39*
RT pumps preceding explosions (ms)	41	738.80 (251.65)	699.53 (234.36)	.27	.58*F	25	.40*

Note.
* significant test-retest reliability/familiality based on 95% quantile of permutations, including control for multiple comparisons.
[a] Number of subjects per variable varies due to the outlier detection/exclusion procedure, see methods for details; (m, SD): mean, standard deviation; RT: reaction time; N: number; ms: millisecond; F: fair ICC values (.4<ICC<.59), G: good ICC values (.6<ICC<.74), based on (Cicchetti, 1994).

2.6. Estimation of motion, temporal signal-to-noise ratio (tSNR), and contrast-to-noise (CNR) ratio

The rotation and translation motion parameters per volume[9] were estimated by the HCP fMRIVolume pipeline (using FSL's MCFLIRT tool).

[6] 80 mm^2 and 120 mm^3 correspond to projection of 20 voxels (2mm × 2mm × 20voxels) on the surface for cortical regions and 15 voxels of volume (2mmX2mmX2mmX15voxels) for subcortical structures, respectively. These values were selected upon visual inspection of cluster extent in activation maps and also taking into account that the cortical areas by definition are larger as compared to subcortical structures.

[7] These effect size thresholds were selected after the inspection of the activation maps in order to select spatially confined regions with the largest effect size. We used different effect size thresholds for cortical and subcortical regions and for the two designs (categorical and parametric) because of large differences in the overall activation magnitude and effect size (cortical greater than subcortical, categorical greater than parametric). Using the same effects size threshold would preclude the identification of discrete regions. For example, applying a threshold that is optimal for discrimination of subcortical regions to the cortical regions would result in clusters of activation spanning >50% of a hemisphere. Conversely, applying thresholds that are optimal for differentiating cortical activation to subcortical regions would have resulted in missing virtually all significantly activated subcortical regions.

[8] Note that all regions that had significant activation present in >50% of the parcel/segment also had Cohen's D > 0.35 for the cortical and >0.2 for the subcortical regions in the categorical design and Cohen's D > 0.2 for the cortical and subcortical regions in the parametric design. Thus, the Cohen's criterion turned out to not have any practical impact on the selection of regions for the *thresholded subset* analysis.

[9] 'prefiltered_func_data_mcf.par' output file.

Table 2

Test-retest reliabilities (TRR ICCs) and familiality (MZ twin correlations) for task related active regions (*thresholded parcels)* and *thresholded subset* (in bold). Only parcels that passed significance testing for the test-retest ICCs are listed[b].

Contrasts	HCP-MP1.0 parcel name	Corresponding Desikan-Killiany Atlas	TRR ICC	Familiality
CATEGORICAL DESIGN				
ChooseInflate	R_VMV3	fusiform	.66	.58
Thresholded Parcels	R_FFC	fusiform	.66	.66[a]
Cutoff TRR ICC = .49	L_VVC	fusiform	.65	.52
Cutoff familiality = .62	R_LO1	lateraloccipital	.64	.28
	L_FFC	fusiform	.62	.55
Thresholded Subset	L_PIT	fusiform, lateraloccipital	.62	.41
Cutoff TRR ICC = .41	R_V4	lateraloccipital, fusiform, lingual	.61	.55
Cutoff familiality = .52	L_V2	lingual, lateraloccipital, cuneus	.60	.45
	L_IPS1	**superiorparietal**	**.59**	**.16**
	L_V3B	inferiorparietal, superiorparietal, lateraloccipital	.58	.31
	R_PH	**lateral occipital, fusiform, inf. temp.**	**.57**	**.59**[a]
	R_PIT	fusiform, lateraloccipital	.57	.39
	R_VVC	fusiform	.57	.56
	R_8BM	superiorfrontal	.56	.33
	R_PGp	inferiorparietal	.56	.58
	R_LO2	lateraloccipital	.56	.34
	R_V3B	inferiorparietal, superiorparietal, lateraloccipital	.55	.50
	L_LIPd	**superiorparietal**	**.55**	**.41**
	L_V4	lateraloccipital, fusiform, lingual	.55	.52
	L_V3	superiorparietal, lateraloccipital, lingual	.55	.55
	L_AAIC	insula	.54	.22
	L_TE2P	inferiortemporal	.54	.62[a]
	L_LO2	lateraloccipital	.53	.36
	R_6a	superiorfrontal/caudal middlefrontal	.52	.38
	R_PFm	Inferiorparietal, supramarginal	.51	.37
	R_V3	superiorparietal, lateraloccipital, lingual	.50	.46
	L_7 PC	superiorparietal	.50	.30
	L_SCEF	superiorfrontal	.50	.27
	R_p32pr	**superiorfrontal**	**.50**	**.11**
	R_46	rostralmiddlefrontal	.49	.35
	R_MIP	**superiorparietal**	**.50**	**.46**
	L_IP1	**inferiorparietal**	**.44**	**.44**
	R_IP1	**inferiorparietal**	**.43**	**.54**[a]
ChooseCashout	**L_AIP**	**superiorparietal, supramarginal**	**.72**	**.69**[a]
Thresholded Parcels	**R_IP0**	**inferiorparietal**	**.67**	**.39**
Cutoff TRR ICC = .52	R_23d	posterior cingulate	.66	.43
Cutoff familiality = .66	L_RSC	posterior cingulate, isthmuscingulate	.66	.34
	L_MIP	superiorparietal	.65	.66[a]
Thresholded Subset	**L_PFt**	**supramarginal, postcentral**	**.64**	**.39**
Cutoff TRR ICC = .44	R_V2	lingual, lateraloccipital, cuneus	.63	.39
Cutoff familiality = .55	R_PHT	middletemporal, inferiortemporal	.63	.50
	R_V8	fusiform	.63	.37
	R_IP1	**inferiorparietal**	**.63**	**.40**
	R_V4	lateraloccipital, fusiform, lingual	.62	.49
	R_V1	lateraloccipital, lingual, pericalcarine, cuneus	.62	.48
	R_LO1	lateraloccipital	.62	.28
	L_POS2	Precuneus, superiorparietal	.61	.44
	R_IPS1	superiorparietal	.61	.16
	R_PIT	fusiform, lateraloccipital	.60	.40
	L_DVT	Superiorparietal, precuneus	.60	.12
	L_IP0	inferiorparietal	.58	.39
	L_7 PL	superiorparietal	.57	.50
	L_LIPv	superiorparietal	.57	.56
	L_IPS1	**superiorparietal**	**.57**	**.44**
	R_V3	superiorparietal, lateraloccipital, lingual	.56	.38
	L_TE2p	inferiortemporal	.56	.34
	R_IFSa	**rostralmiddlefrontal, parstriangularis**	**.56**	**.66**[a]
	R_PH	**lateral occipital, fusiform, inf. temp.**	**.56**	**.34**
	R_TE1p	Middletemporal, inferiortemporal	.56	.41
	R_RSC	posterior cingulate, isthmuscingulate	.55	.13
	R_31a	Posterior cingulate	.55	.25
	R_FST	middletemporal, lateraloccipital, inf. temp.	.55	.49
	R_PFm	**inferior parietal, supramarginal**	**.55**	**.28**
	R_VVC	fusiform	.55	.50
	L_PH	lateral occipital, fusiform, inf. temp.	.54	.36
	R_d23ab	Posteriorcingulate, istmuscingulate	.54	.56
	L_IP2	supramarginal, inferiorparietal	.52	.43
	L_PGs	inferiorparietal	.52	.41
	L_TE2a	Inferior/middle temporal	.52	-.14
	R_MIP	**superiorparietal**	**.48**	**.34**
	L_IP1	**inferiorparietal**	**.47**	**.30**
	R_d32	**superior frontal, rostral anterior cingulate**	**.47**	**.31**
	L_a9_46v	**rostral middle frontal**	**.46**	**.45**
	R_AIP	**superiorparietal, supramarginal**	**.44**	**.28**
OutcomeInflate	R_V3	superiorparietal, lateraloccipital, lingual	.56	.53
Thresholded Parcels	R_V1	lateraloccipital, lingual, pericalcarine, cuneus	.54	.48
Cutoff TRR ICC = .47	R_V2	lingual, lateraloccipital, cuneus	.53	.47
Cutoff familiality = .59	R_8C	rostral/caudal middle frontal	.50	.25
	R_V4	lateraloccipital, fusiform, lingual	.49	.37
Thresholded Subset	L_V2	lateraloccipital, lingual, cuneus	.49	.46
Cutoff TRR ICC = .33				
Cutoff familiality = .43				
OutcomeExplode	L_V3B	inferiorparietal, superiorparietal, lateraloccipital	.64	.21
Thresholded Parcels	R_LO1	lateraloccipital	.57	.33
	R_TE2p	inferiortemporal	.53	.13

(*continued on next page*)

Table 2 (*continued*)

Contrasts	HCP-MP1.0 parcel name	Corresponding Desikan-Killiany Atlas	TRR ICC	Familiality
Cutoff TRR ICC = .52				
Cutoff familiality = .65	L_9a	superior frontal, rostral middle frontal	.53	-.01
	L_PIT	fusiform, lateraloccipital	.52	.31
Thresholded Subset	**L_PGp**	**inferiorparietal**	**.47**	**.27**
Cutoff TRR ICC = .45				
Cutoff familiality = .56				
PARAMETRIC DESIGN				
ChooseCashout	L_FST	middletemporal, lateraloccipital, inf. temp.	.53	-.10
Thresholded Parcels	L_IPS1	superiorparietal	.52	.47
Cutoff TRR ICC = .47	R_V1	lateraloccipital, lingual, pericalcarine, cuneus	.49	.34
Cutoff familiality = .62	L_IP0	inferiorparietal	.48	.52
	R_MIP	**superiorparietal**	**.45**	**.45**[a]
Thresholded Subset				
Cutoff TRR ICC = .15				
Cutoff familiality = .32				
OutcomeInflate	R_PIT	fusiform, lateraloccipital	.51	.14
Thresholded Parcels				
Cutoff TRR ICC = .45				
Cutoff familiality = .58				
Thresholded Subset				
Cutoff TRR ICC = .22				
Cutoff familiality = .42				
OutcomeExplode	**L_PGp**	**inferiorparietal**	**.56**	**-.02**
Thresholded Parcels	**R_PH**	**lateral occipital, fusiform, inf. temp.**	**.33**	**-.03**
Cutoff TRR ICC = .54				
Cutoff familiality = .68				
Thresholded Subset				
Cutoff TRR ICC = .30				
Cutoff familiality = .55				

Notes. n for the twin correlations ranged from 19 to 27, due to outlier detection procedure (see *Outlier Detection and Exclusion* section).

[a] Regions with significant test-retest reliability and MZ twin correlations, which are good candidate regions for phenotypes for future genetic studies. For the *thresholded subset* analysis, the permutation-based cutoffs were lower than for the *thresholded parcels* analysis due to fewer total regions, and thus a less severe correction for multiple comparisons.

[b] Note that among *thresholded parcels*, only in the categorical design, parcels L_V8 (familiality = .67, TRR ICC = .43) during *ChooseInflate*, L_6v (familiality = .66, TRR ICC = .51) and L_FOP2 (familiality = .65, TRR ICC = .52) during *ChooseCashout*, and R_LIPd (familiality = .64, TRR ICC = .16) during *OutcomeExplode* contrasts showed statistically significant familiality but non-significant TRR ICCs.

The average of the frame-to-frame movement for each run[10] was calculated for Time1 and Time2, and then averaged across Time1/Time2 for each person.

In consideration of previous research showing a relationship between temporal SNR and TRRs (Raemaekers et al., 2007), we investigated whether differences in tSNR were able to explain the differences in TRRs across parcels and across categorical and parametric designs and if this was moderated by the level of motion. Temporal SNR for each grayordinate was calculated as the mean over time divided by the square root of the variance estimated from the residuals after model fitting,[11] for Time1. Grayordinate-wise tSNR values were first parcellated (values across grayordinates within each parcel/segment were averaged) and then averaged across subjects (full sample of 56 subjects). Since our data showed that tSNRs for the subcortical regions were significantly smaller than the cortical regions in both designs and at both time-points (all $p < .001$), we present our findings for the *untresholded parcels* separately for the cortical and subcortical regions.

Contrast-to-Noise (CNR) ratios for Time1 were also calculated for each *unthresholded parcel.* Grayordinate-wise variance estimated from the residuals after model fitting was first parcellated (values across grayordinates within each parcel/segment were averaged), then square root of that mean was taken. CNR for a parcel was calculated as the mean beta weights per parcel within a contrast divided by the square root of the variance described above; which provided us with information on sensitivity of each specific contrast separately.

2.7. Outlier detection and exclusion

Each behavioral variable, mean BOLD response magnitude of *unthresholded* and *thresholded parcels*, motion, tSNR and CNR estimates were analyzed for outliers in R Core Team (2018) (https://www.R-project.org/). This procedure was applied to the whole sample, separately on the Time1 (n = 56) and Time2 (n = 44) data. For the outlier detection procedure only, raw values were converted to Z-scores, and then values greater than three standard deviations from zero were recoded as missing values. This procedure was reiterated 10 times as outlier removal changes the shape of the distribution, allowing for the emergence of new outliers. With this exclusion procedure, 1% of the behavioral data (on average from Time1 and Time2 data altogether) was replaced with missing values, 1.71% and 4.42% of the *unthresholded parcels* mean BOLD data, 1.7% and 5.11% of the *thresholded parcels* mean BOLD data, 6.95% and 6.97% of the *unthresholded parcels* variance estimated from the residuals in the categorical and parametric designs, respectively. 7.14% of the motion at Time1, none of the motion at Time2, 2.41% of the tSNR at Time1, and 0.24% of tSNR at Time2 were also replaced with missing values.

2.8. Test-retest reliability estimates

Test-retest reliabilities (TRRs) were estimated for the behavioral measures and the *unthresholded* and *thresholded parcels* data. Although, intraclass correlation (ICC) is one of the most commonly used test-retest reliability measures in the neuroimaging field (Bennett and Miller, 2010), there are several other methods to assess reliability, such as Pearson correlation, coefficient of variation, Cohen's kappa index, and Kendall's W. It is noteworthy that ICC estimates are specific to the dataset under investigation, which limits the generalizability of ICCs estimated in controls to clinical samples. It is important to note that with two time-point "consistency" ICC used in the present study, the ICC values were highly convergent with the Pearson correlation between the two measurement occasions. ICCs are typically calculated as the ratio of the between-subject variance and total variance (Shrout and Fleiss, 1979)

[10] 'Movement_RelativeRMS_mean.txt' output file.

[11] tfMRI_*_Atlas.mean.dscalar.nii/$\sqrt{}$sigmasquareds.dtseries.nii.

Table 3
Correlations between measures of test-retest reliability (**TRR** ICCs), effect sizes (Cohen's d), Time1 activation (beta weights, mean and std across participants), and familiality (MZ twin correlations) for the *unthresholded parcels* analysis (i.e., whole brain parcellation/segmentation)..

		CORTICAL MMP PARCELLATION						SUBCORTICAL FREESURFER SEGMENTATION					
CATEGORICAL DESIGN	**Choose Inflate**		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**
		ICC	1	.46**	.51**	.38**	.40**	**ICC**	1	.29	.08	-.52	-.09
		CohensD		1	.94**	.31**	.23**	**CohensD**		1	.93**	-.09	.24
		MBetas			1	.54**	.25**	**MBetas**			1	.28	.31
		SDBetas				1	.28**	**SDBetas**				1	.43
		Familiality					1	**Familiality**					1
	Choose Cashout		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**
		ICC	1	.60**	.60**	.42**	.46**	**ICC**	1	.15	.47*	.68*	.61**
		CohensD		1	.94**	.42**	.52**	**CohensD**		1	.88**	.29	.48*
		MBetas			1	.66**	.54**	**MBetas**			1	.70**	.69**
		SDBetas				1	.35**	**SDBetas**				1	.64**
		Familiality					1	**Familiality**					1
	Outcome Inflate		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**
		ICC	1	.01	.07	.22**	.19**	**ICC**	1	-.005	-.002	.04	-.21
		CohensD		1	.97**	.12*	.19**	**CohensD**		1	.92**	.28	.30
		MBetas			1	.23**	.19**	**MBetas**			1	.63**	.27
		SDBetas				1	.10	**SDBetas**				1	.01
		Familiality					1	**Familiality**					1
	Outcome Explode		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**
		ICC	1	.40**	.48**	.40**	.16**	**ICC**	1	-.18*	-.12	.37	.16
		CohensD		1	.92**	.13*	.06	**CohensD**		1	.95**	-.29	.01
		MBetas			1	.42**	.15**	**MBetas**			1	-.003	.06
		SDBetas				1	.23**	**SDBetas**				1	.31
		Familiality					1	**Familiality**					1
PARAMETRIC DESIGN	**Choose Inflate**		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**
		ICC	1	-.08	-.07	.02	-.04	**ICC**	1	.13	.18	-.42	.16
		CohensD		1	.98**	-.01	-.02	**CohensD**		1	.98**	-.11	-.08
		MBetas			1	-.15**	-.02	**MBetas**			1	-.19	-.14
		SDBetas				1	.06	**SDBetas**				1	.47*
		Familiality					1	**Familiality**					1
	Choose Cashout		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**
		ICC	1	.41**	.41**	.13*	.31**	**ICC**	1	.12	.31	.47*	.54*
		CohensD		1	.88**	.16**	.46**	**CohensD**		1	.89**	-.09	.21
		MBetas			1	.54**	.41**	**MBetas**			1	.33	.20
		SDBetas				1	.03	**SDBetas**				1	.20
		Familiality					1	**Familiality**					1
	Outcome Inflate		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**
		ICC	1	.28**	.27**	.04	.25**	**ICC**	1	-.12	-.19	-.33	.24
		CohensD		1	.97**	.03	.18**	**CohensD**		1	.89**	-.23	-.06
		MBetas			1	.16**	.19**	**MBetas**			1	.19	.08
		SDBetas				1	-.03	**SDBetas**				1	.10
		Familiality					1	**Familiality**					1
	Outcome Explode		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**		**ICC**	**CohensD**	**MBetas**	**SDBetas**	**Familiality**
		ICC	1	.38**	.40**	.04	-.04	**ICC**	1	.17	.07	-.24	.19
		CohensD		1	.86**	-.26**	.01	**CohensD**		1	.83**	-.54*	.61**
		MBetas			1	.19**	.04	**MBetas**			1	-.14	.59**
		SDBetas				1	.06	**SDBetas**				1	-.15
		Familiality					1	**Familiality**					1

Notes. ICC: Intraclass correlation coefficients; CohensD: effect sizes based on Time 1 data (mean/SD of beta weights across subjects -based on full sample except outliers-, calculated per parcel); MBetas: mean of Time 1 beta weights; SDBetas: standard deviation of Time 1 beta weights; Familiality: correlations of beta weights between monozygotic twins. The FDR corrected significant clusters were further divided into anatomical parcels using the HCP-MMP1.0 (Human Connectome Project Multi-Modal Parcellation version 1.0) cortical parcellation (Glasser et al., 2016) and Freesurfer subcortical segmentation (Fischl et al., 2002). Pearson correlations, *< 0.05, **<0.01. Also see Supplementary Figure S21, for scatterplots depicting the correlations between mean beta weights at Time1 and ICCs across cortical parcels and subcortical segments together.

and represent the stability of individual differences in the degree of activation over time. In this study, reliability was quantified as the degree of consistency between the Time1 and Time2 measurements, under the assumption of a two-way mixed model, which is known as ICC(3,1) (Shrout and Fleiss, 1979), or alternatively ICC(C,1) (McGraw and Wong, 1996). The relevant mean squares were estimated using method of moments estimators and a Matlab function ('ICC.m'[12]) based on ICC(C,1) with the use of formulas provided by (McGraw and Wong, 1996) (referred as 'TRR ICC' in text). Note that this estimator allows for negative ICCs, which were retained in the data to maintain the overall distribution of reliabilities.

Cicchetti (1994) proposed that ICCs are considered poor, fair, good, and excellent when ICC<0.4, 0.4<ICC<0.59, 0.6<ICC<0.74, 0.75<ICC<1, respectively. The ICC estimates obtained from a study are only an expected value of the true ICCs. Koo and Li (2016) recommended not only to decide the degree of reliability (poor, fair, good etc.), but also to determine the reliability of ICC estimates themselves using statistical inference, their suggested methods were mainly based on parametric assumptions. Termenon et al. (2016) instead recommended the use of permutation tests rather than parametric assumptions (F-test) to determine whether ICCs are significantly different from zero, since the use of parametric approaches for assessing the significance of ICCs can be too restrictive with small samples or when the sample distribution does not conform with parametric assumptions (e.g., independence and gaussianity). Therefore in this study, we implemented a permutation method (5000 permutations) to determine the significance of ICC estimates, which also provided a convenient mechanism to control for the testing of multiple hypotheses (i.e., across all behavioral variables and across all parcels/segments). In this procedure, the Time2 data were randomly permuted – i.e., relabeled as the Time2 for a different participant (without regard to twinship) – and ICCs were re-calculated for each of the 5000 permutations. A null distribution was created by selecting the highest ICC (across behavioral measures or parcels/segments) in each permutation. ICC values greater than or equal to the 95th quantile of this null distribution were considered as statistically significant. Permutations were completed separately for the behavioral variables, *unthresholded parcels*, *thresholded parcels* and for the *thresholded subset* (for the distribution of permuted ICCs, see *Supplementary Materials*, Figures S5, S6, S7 and S8). Only parcels with significant ICCs are reported in this paper; however ICCs of the full *unthresholded parcels*, and *thresholded parcels* can be found in the BALSA repository for neuroimaging data: https://balsa.wustl.edu/study/show/87PP9.

2.9. Familial influences

Since the present sample included MZ twins, we examined the degree of twin resemblance (intrapair twin correlations) with respect to the magnitude of task-related activations and compared it with the test-retest reliability estimates. MZ twin correlation represents a direct measure of familiality of individual differences (familial transmission) that includes genetic factors (both additive and non-additive) and shared environment. We expected that more reliable regions would also show higher degree of familiality, since familiality is inherently bounded by reliability.

Estimates of twin correlations may be subject to bias due to the randomness of assigning twin pairs to "twin 1" and "twin 2". Contrary to some alternative methods of quantifying twin correlations, ICC values are not affected by this random assignment, therefore twin correlations in this study were calculated using the same ICC(C,1) estimator used for the TRR estimation (see section, *Test-Retest Reliability Estimates*). This estimate of 'familiality' was based on the Time1 data (n = 27 pairs), which allowed us to estimate twin correlations with the maximum number of twin pairs. The same permutation approach that was utilized for the significance testing of the ICCs (see section, *Test-Retest Reliability Estimates*) was also applied for the significance testing of the twin correlations, except in this case, the sibship assignment of "twin 2" was randomly permuted. Twin correlations greater than or equal to the 95th quantile of permuted null distribution were considered as significant. Permutations were completed separately for the behavioral variables, *unthresholded parcels*, *thresholded parcels* and for the *thresholded subset* (for distribution of permuted twin correlations, see *Supplementary Materials*, Figures S5, S9, S10, and S11).

3. Results

3.1. Performance results

Mean values of behavioral outcome variables measured with the in-scanner BART did not change significantly over time, except that the reaction time for cashing out (risk averse decision) decreased from Time1 to Time2 (see Table 1). Overall, between session test-retest reliability for behavioral measures was moderate-to-high (ICCs > 0.40). ICC correlations were slightly higher for the outcome variables of cash-outs compared to the outcome variables of balloon explosions. The main behavioral outcome measure of risk-taking (average number of pumps before cash-outs) showed a good test-retest reliability (ICC = 0.67, p < .01). However, with the exception of a few reaction time measures, familial correlations were generally low, even for variables showing high reliability.

3.2. fMRI results

3.2.1. Test-retest reliability of unthresholded parcels, thresholded parcels, and thresholded subset

Fig. 2 displays the ICC values for the *unthresholded parcels* with TRR ICCs > 0.2. Moreover, Supplementary Table S3 lists cortical parcels and subcortical segments with significant ICCs for all task conditions and both designs for the same *unthresholded parcels*.

ICC values tended to be greater in the categorical design compared to the parametric design (see *Supplementary Materials* Figure S12). Moreover, ICC values tended to be greater for the cortical parcels compared to the subcortical segments. Within the categorical design, the *Choose-Cashout* contrast tended to have the largest ICC values (mean TRR ICC = 0.25), followed by the *ChooseInflate* (mean TRR ICC = 0.22), *OutcomeInflate* (mean TRR ICC = 0.11) and *OutcomeExplode* (mean TRR ICC = 0.11) contrasts. Within the parametric design, *OutcomeInflate* contrast tended to have the largest ICC values (mean TRR ICC = 0.16), followed by *ChooseCashout* (mean TRR ICC = 0.14), *OutcomeInflate* (mean TRR ICC = 0.13) and *OutcomeExplode* (mean TRR ICC = 0.08) contrasts.

Significant ICC values across parcels ranged from 0.51 to 0.77 and from 0.52 to 0.60 for the categorical and parametric designs, respectively. Regions with the largest ICCs included the superior/inferior parietal area (range TRR ICCs = 0.54-0.77), lateral occipital area (range TRR ICCs = 0.52-0.64), the fusiform gyrus (range TRR ICCs = 0.54-0.64), and superior/inferior temporal regions (range TRR ICCs = 0.54-0.60) in all of the contrasts. Within the subcortical regions, the right caudate had the highest test-retest reliability (TRR ICC = 0.63). Lastly, in order to investigate a potential bias in reliability estimates due to dependencies in the data (MZ twin correlations), reliabilities of the *unthresholded parcels* were re-estimated for two unrelated groups, by assigning Twin 1 and Twin 2 of our twin pairs to separate groups. This analysis revealed that reliabilities of the *unthresholded parcels* in the average of the two subsamples of unrelated individuals were highly correlated with those derived from the full sample (all r's > 0.96, all p's < 0.001, see *Supplementary Materials*, Fig. S20).

Table 2 lists the regions with significant ICCs among all task-related active parcels (*thresholded parcels*) and among the *thresholded subset* (also see Fig. 3). Overall, areas activated during risky and risk-averse

[12] https://www.mathworks.com/matlabcentral/fileexchange/22099-intraclass-correlation-coefficient-icc.

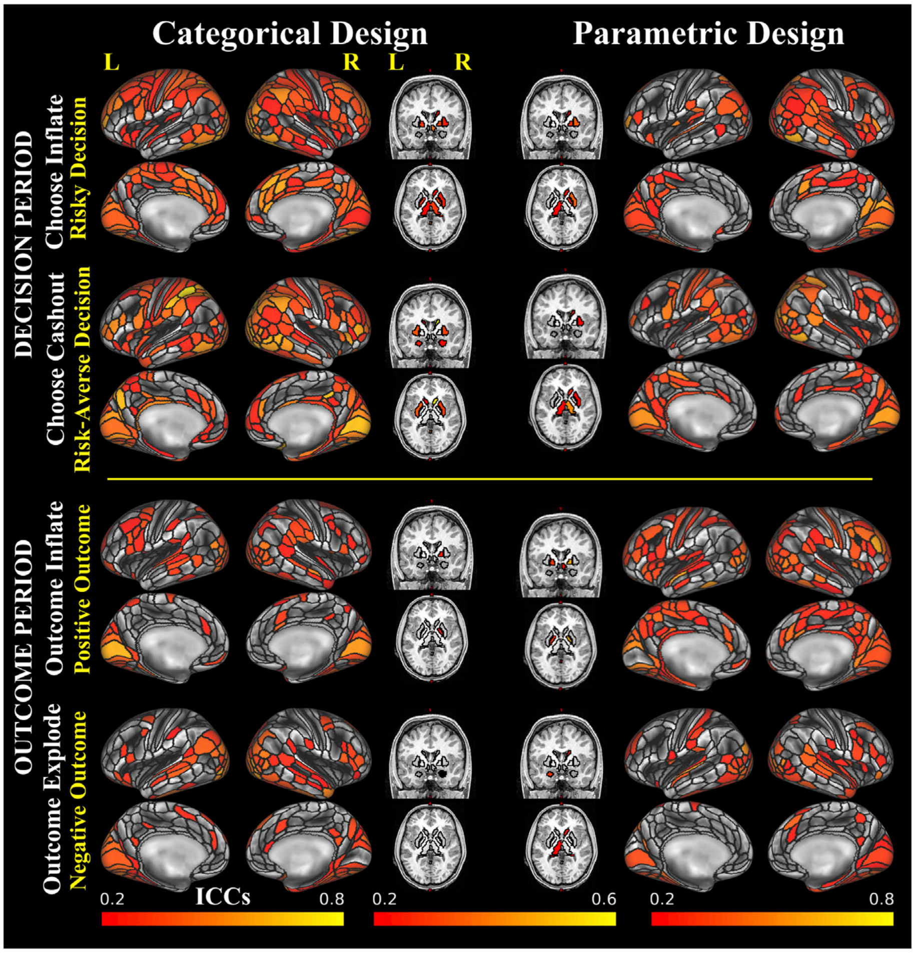


Fig. 2. Test-retest reliabilities (ICCs) for the whole brain MMP cortical parcels and Freesurfer subcortical segments (*unthresholded parcels* analysis) varied from none-to-high. ICCs are plotted separately for the categorical and parametric designs during the decision making and outcome (feedback) phases of the BART. Regions with ICC<.2 ("low" correlation) are of little interest and are masked by gray color. ICCs for all cortical and subcortical regions were lower than 0.8 and 0.6, respectively. L: left, R: right.

decisions tended to have the largest ICCs.

Within broader task related active regions (*thresholded parcels)*, lingual, inferior temporal, fusiform, lateral occipital, superior and inferior parietal areas had the largest ICCs during both risky decisions (*ChooseInflate*) and risk-averse decisions (*ChooseCashout*). The superior frontal and insula were the other two regions with high reliabilities in the risky decision condition. Rostral middle frontal, posterior cingulate, precuneus, middle temporal and supramarginal areas had high reliabilities during risk-averse decisions.

Within the *thresholded subset*, in the categorical design, superior parietal, inferior parietal, lateral occipital and fusiform showed the highest TRRs during risk taking (*ChooseInflate*). Superior and inferior parietal, lateral occipital, rostral middle frontal, and pars triangularis areas showed the highest reliabilities during risk-averse decisions (*ChooseCashout*) and inferior parietal during evaluation of negative outcomes (*OutcomeExplode)*. In the parametric design, superior parietal (*ChooseCashout*) and inferior parietal/lateral occipital/fusiform (*OutcomeExplode*) areas showed the highest reliabilities. No other regions were identified as significantly reliable in the other contrasts of the parametric design in the *thresholded subset* analysis.

3.2.2. *Familiality and its relationship with test-retest reliability (unthresholded parcels)*

Fig. 4 shows the relationship between TRR (ICCs) and familiality (MZ twin correlations) at Time1 for the *unthresholded parcels*. The spatial distribution of the MZ twin correlations is also presented in the Supplementary Figure S13. Across all parcels, familiality ranged from −0.60 to 0.80 and tended to be larger in the *ChooseInflate* and *ChooseCashout* contrasts of the categorical design. Regions with both significant ICCs and significant familiality were identified in the *ChooseInflate* (risky decision) and *ChooseCashout* (risk-averse decision) contrasts with the use of categorical design, and included the right fusiform, right rostral anterior cingulate (ACC), and left superior parietal regions.

Table 3 provides an overall summary of correlations between measures of test-retest reliability and familiality (MZ twin correlation), separately for the cortical and subcortical regions. These correlations show the extent to which reliability predicts familiality. In the categorical design, across the cortical parcels, familiality correlated with the measures of test-retest reliability moderately in the *ChooseInflate* ($r = 0.40$, $p < .01$) and *ChooseCashout* ($r = 0.46$, $p < .01$) contrasts and weakly in the *OutcomeInflate* ($r = 0.19$, $p < .01$) and *OutcomeExplode* ($r = 0.16$, $p < .01$) contrasts. For the subcortical regions, familiality was correlated with test-retest reliability only in the *ChooseCashout* contrast ($r = 0.61$, $p < .01$) for the categorical design. In the parametric design, familiality showed weak-to-moderate across-parcels correlations with the measures of test-retest reliability in the cortical parcels in the *ChooseCashout* ($r = 0.31$, $p < .01$) and *OutcomeInflate* ($r = 0.25$, $p < .01$) contrasts only. In subcortical regions, familiality moderately correlated with the test-retest reliability in the *ChooseCashout* contrast ($r = 0.54$, $p < .05$) of the parametric design.

3.2.3. *Factors potentially affecting test-retest reliability*

3.2.3.1. Does the magnitude of activation predict test-retest reliability?. To address this question, we computed correlations across parcels/segments of the effect size (Cohen's d) and magnitude (mean beta) of unthresholded activation within a parcel with the TRR of the mean beta weights within that parcel (Table 3). In the categorical design, and across the

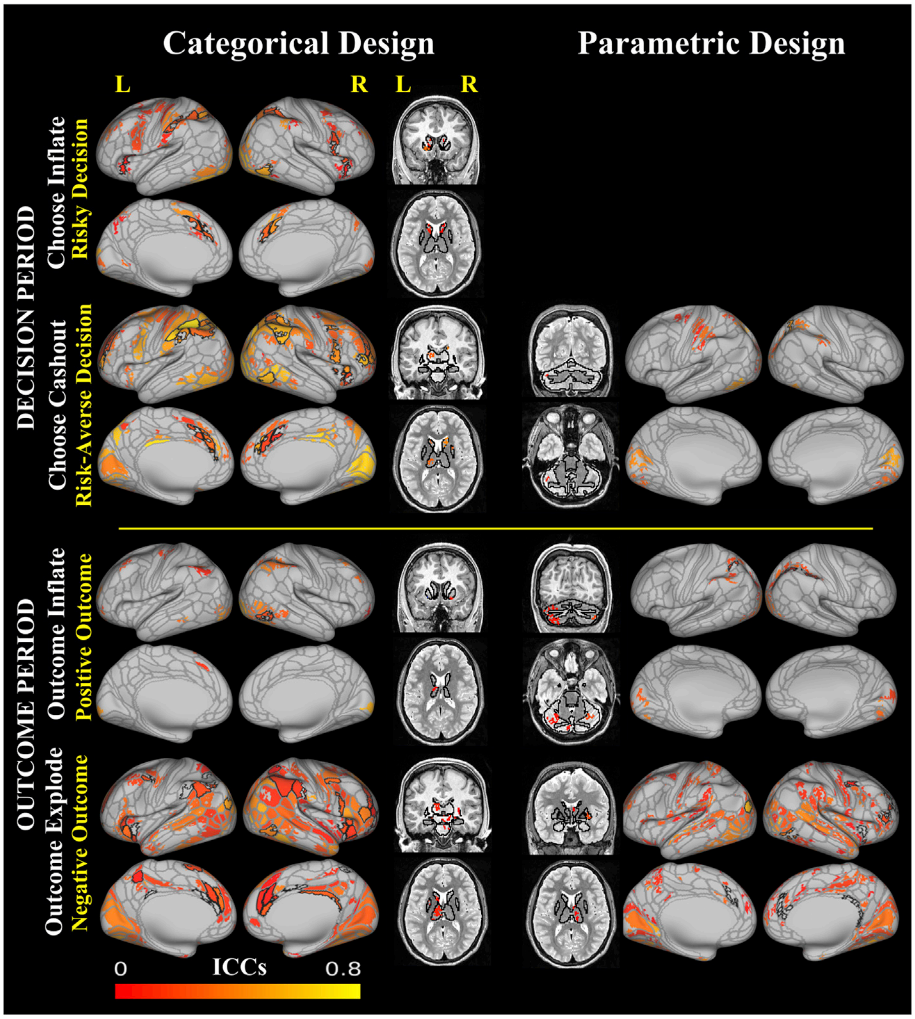


Fig. 3. Test-retest reliabilities (ICCs) of significant task-related activations varied from none-to-high across different parcels. The FDR corrected significant clusters were further divided into anatomical segments using the Human Connectome Project Multi-Modal Parcellation (MMP1.0) and the Freesurfer subcortical segmentation. ICCs were mapped separately for the decision making and outcome (feedback) phases of the BART for the categorical and the parametric designs. On cortical surface view, *Gray* outlines depict the boundaries of the MMP1.0 cortical parcellation and *Black* outlines depict the *thresholded subset*. Cortical parcels with negative ICC values are not plotted. On subcortical volume view, *Black* outlines depict the Freesurfer segmentation. All cortical and subcortical ICCs were lower than 0.8. L: left, R: right.

cortical parcels, effect size and magnitude of the Time1 activations showed moderate-to-strong positive correlations (ranging from r = 0.40 to 0.60) with the test-retest reliability for all contrasts (except for the *OutcomeInflate* contrast; i.e., outcome evaluation of risky decisions). Across the subcortical segments the effect size and the magnitude of Time1 activations were not related to the measures of test-retest reliability, except a moderate correlation between the magnitude of Time1 activations in the *ChooseCashout* contrast (r = 0.47). In the parametric design, the effect size and magnitude of the Time1 cortical activations predicted test-retest reliability across the cortical parcels in all contrasts (range of r = 0.27 to 0.41), except for the *ChooseInflate* contrast (risk-taking). In subcortical regions, there were no significant correlations between the TRR and the effect size or magnitude of Time1 activations for the parametric design. Overall as the magnitude of beta weights (unthresholded activations) increases, the TRR increases as well; however this affect is mostly driven by cortical regions. Moreover parcels with greater magnitude of activation also had greater variability, which would result in greater between subject individual differences in these parcels. With increasing beta weights, we also observed greater tSNRs (see Supplementary Figure S17), and subcortical regions were at lower end of the spectrum on both beta and tSNR measures, which explains the observed low reliabilities in subcortical structures.

3.2.3.2. Do 'thresholded parcels' show greater test-retest reliability than regions defined without regard to activation?. Fig. 5 depicts the relationship between test-retest reliability for the *thresholded parcels* and *unthresholded parcels.* Overall, ICCs for the *thresholded parcels* were significantly greater than the *unthresholded parcels* in the categorical design only ($p < .001$). However, ICC values for these two approaches to defining a region correlated strongly in both designs (r = 0.77 in the categorical and r = 0.62 in the parametric design), suggesting that regions with high and low TRRs were largely consistent across different ways of defining the boundary of a region. However, it is relevant to note that while some regions had higher reliability when their boundaries were defined without regard to activation (*unthresholded parcels*), some others had greater reliability when they were defined based on the overlap of the thresholded activation maps and a parcellation scheme (*thresholded parcels*). Per contrast correlations between *thresholded* and *unthresholded parcels* can be found in Fig. 5.

3.2.3.3. How does in-scanner motion affect the consistency of estimated activation?. Since test-retest reliability (ICCs) is computed across a set of participants, it is not possible to correlate reliability with per-participant motion. Instead, we examined the relationship between the average amount of motion across the two sessions (Time1 and Time2) for each participant and the disparity in beta weights from Time1 to Time2 for each participant (quantified as the absolute difference in beta weights across sessions) averaged over all *unthresholded parcels* (Fig. 6). This analysis revealed a moderate correlation between motion and inconsistency of regional activation across sessions (ranging from 0.25 to 0.53 across contrasts and designs, all *ps* < .05, except for the *ChoseInflate* contrast: r = 0.01, p = .99 and *OutcomeInflate* contrast: r = 0.25, p = .1 in the categorical design), suggesting that a greater amount of average motion predicts larger within-subject disparity across two sessions, i.e.,

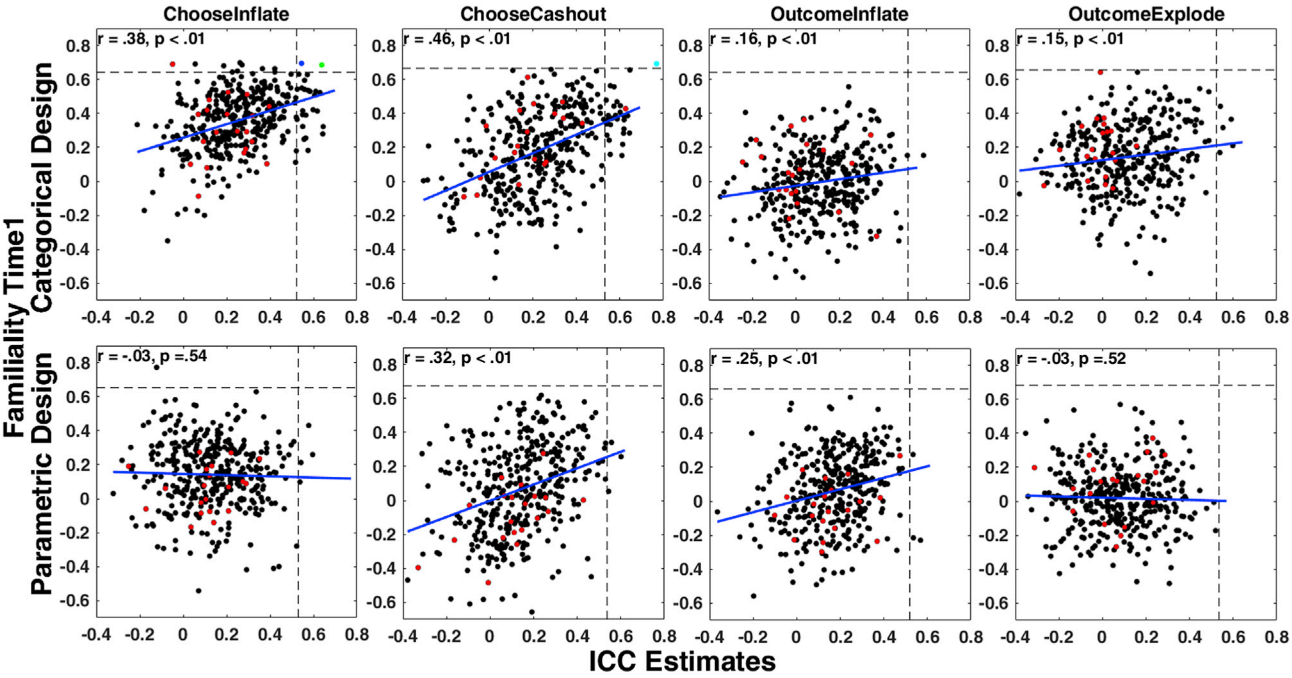
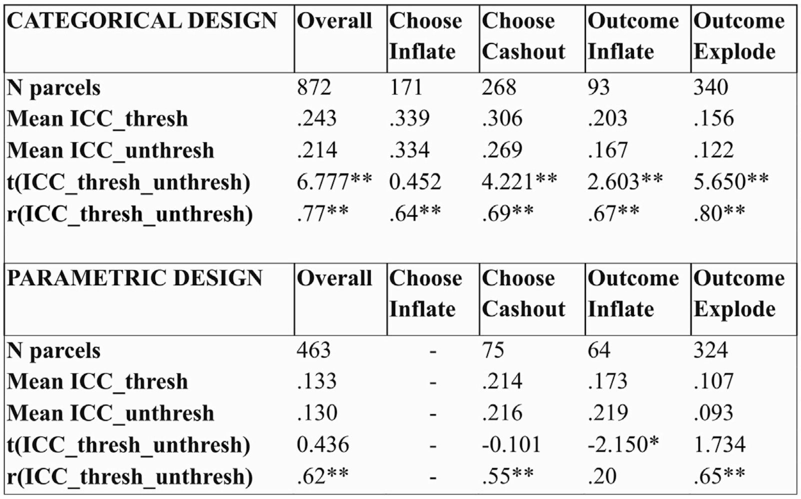


Fig. 4. Test-retest reliability moderately predicts Familiality during risky decisions (*ChooseInflate)* and risk-averse decisions (*ChooseCashout)*. Moreover, reliable and heritable regions were identified in the same contrasts. Scatterplots display test-retest reliabilities (ICCs) and familiality (MZ twin correlations) for the whole brain cortical MMP parcels and subcortical brain regions (*unthresholded parcels* analysis) for categorical and parametric designs during the decision making and outcome (feedback) phases of the BART. Note that each data-point represents a parcel/segment. Regression lines and correlations are calculated based on the joined cortical and subcortical data. Black: Cortical parcels; Red: Subcortical segments; Blue: R-p24 = Rostral ACC; Green: R_FFC = Fusiform area; Cyan: L_MIP = superior parietal area. Dashed lines mark the significance cutoff (95% quantile of permuted null distributions) for the TRR ICC and familiality.



CATEGORICAL DESIGN	Overall	Choose Inflate	Choose Cashout	Outcome Inflate	Outcome Explode
N parcels	872	171	268	93	340
Mean ICC_thresh	.243	.339	.306	.203	.156
Mean ICC_unthresh	.214	.334	.269	.167	.122
t(ICC_thresh_unthresh)	6.777**	0.452	4.221**	2.603**	5.650**
r(ICC_thresh_unthresh)	.77**	.64**	.69**	.67**	.80**

PARAMETRIC DESIGN	Overall	Choose Inflate	Choose Cashout	Outcome Inflate	Outcome Explode
N parcels	463	-	75	64	324
Mean ICC_thresh	.133	-	.214	.173	.107
Mean ICC_unthresh	.130	-	.216	.219	.093
t(ICC_thresh_unthresh)	0.436	-	-0.101	-2.150*	1.734
r(ICC_thresh_unthresh)	.62**	-	.55**	.20	.65**

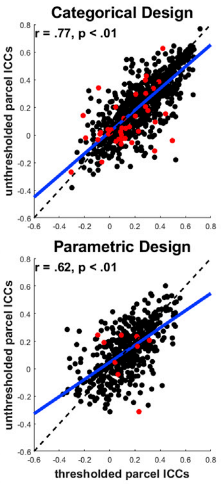


Fig. 5. Thresholded parcels show greater test-retest reliability than regions defined by cortical parcellation and subcortical segmentation schemes only. Relationships between ICC values for the *thresholded parcels* and *unthresholded parcels* analyses for all significant task-related activations across all contrasts in the categorical (n = 872 parcels) and parametric (n = 463 parcels) designs (right panel).

Note. ICC_thresh: ICC values for the average beta values of significant grayordinates within a parcel, i.e., the overlap between a parcel/segment and the significant clusters defined from the whole-brain graydordinate-wise FDR corrected maps (*thresholded parcels* analysis); ICC_unthresh: ICC values for the average beta values of all the grayordinates in the same parcel/ segment (*unthresholded parcels* analysis); *p < .05, **p < .01.

lower within-subject consistency.

Next, we examined whether this moderate relationship between motion and inconsistency in activation beta weights across two sessions at the whole-brain level would manifest for specific parcels among the selected ROIs based on the prior literature as well. We examined the influence of average motion on regions with moderate TRR (fusiform and superior parietal, ICCs = .36 to .59) and low TRR (insula, ACC, ICCs = 0.02 to 0.36) for the *ChooseInflate*, *ChooseCashout*, and *OutcomeExplode* contrasts in the categorical design, as an example. (Activations in these regions did not survive thresholding in the *OutcomeInflate* contrast only). Fig. 7 suggests that although the degree of motion moderately affects the consistency of activation on a broad scale (i.e., when averaged across all parcels), the degree of this influence on individual parcels seems to be lower, and with no obvious difference between regions with low or high reliabilities.

3.2.3.4. Does the tSNR affect test-retest reliability?

Next, we investigated if differences in TRR (ICC) across designs or between cortical and subcortical regions were related to differences in tSNR per parcel. Figure S14 displays the scatterplots for the ICC values and the average tSNR (across participants) at Time1 for all *unthresholded parcels*. Although there was a significant relationship between ICC and tSNR (r ranging

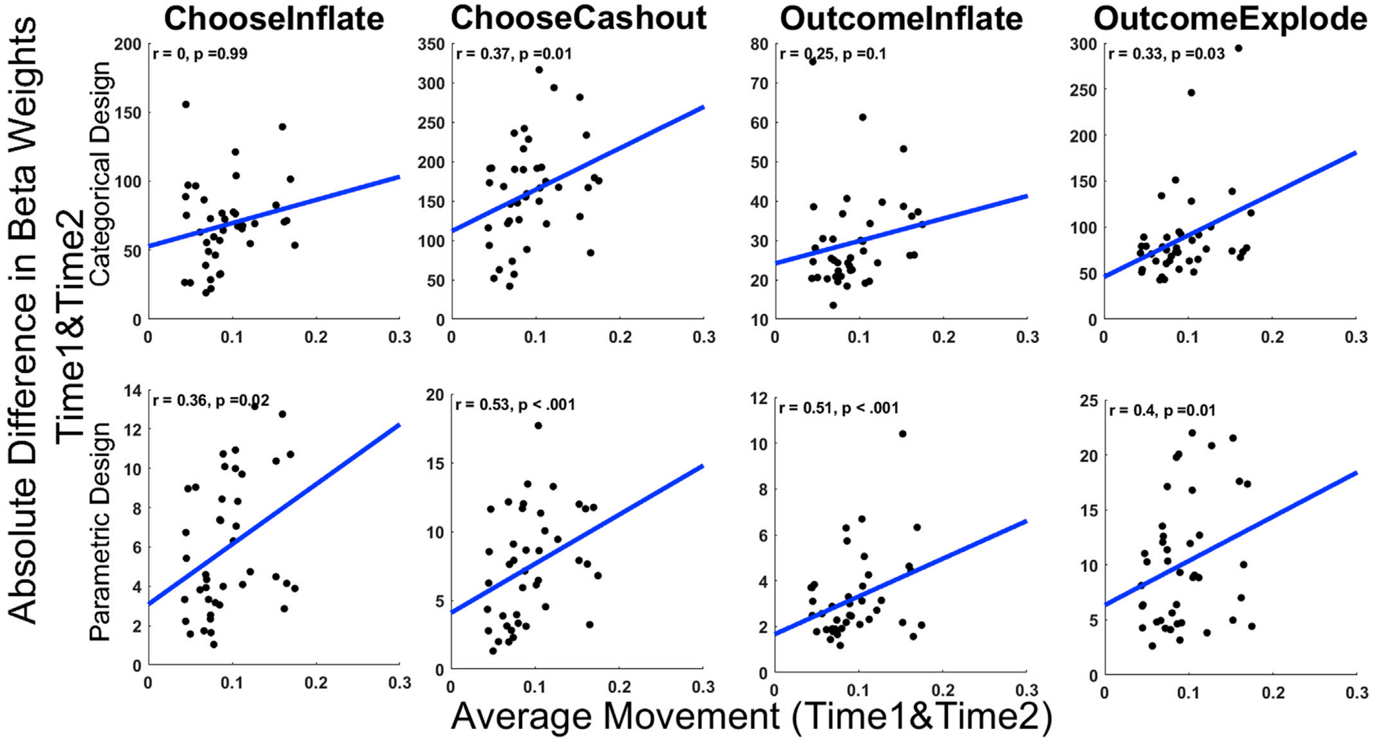


Fig. 6. In-scanner movement predicts intra-individual variability of activation averaged across whole parcels across sessions. Scatterplots of average movement (across run and Time1 and Time2) and disparity in beta weights (absolute difference across Time1 and Time2, from the *unthresholded parcels* analysis) for all contrasts in the categorical and parametric design of BART. Each data-point represents a participant.
Note. Units of movement (mm).

from 0.10 to 0.39, all p's =< 0.05), this relationship was present in all contrasts and in both the categorical and parametric designs. Interestingly, the correlation between ICC and tSNR at Time1 disappeared after regressing motion at Time1 from tSNR at Time1 (see Supplementary Figure S15), indicating that the effect of tSNR on the reliability estimates are in part moderated by the motion. Any effort that would decrease the amount of motion in the data would increase test-retest reliabilities partially by increasing the tSNR in the data.

4. Discussion

The aims of this study were to estimate test-retest reliability of neural correlates of decision making under risk, identify regions with high reliability and familiality of individual differences that can be used as candidate regions (endophenotypes) in clinical and genetic studies, and to examine the factors potentially affecting test-retest reliabilities. In our results, the most important concept is the reliability of task related activations (*thresholded parcels* analysis) and selected regions (*thresholded subset*). Therefore, we discuss those findings first (see section 4.2), followed by the discussion of whole-brain parcellated results (*unthresholded parcels* analysis, see sections 4.3 and 4.4). We have identified reliable regional activations related to risky decisions and positive outcome of risky decision, including bilateral fusiform, bilateral rostral middle frontal, bilateral superior parietal, right lateral occipital, right rostral ACC, left inferior parietal, left caudal ACC, and left inferior temporal regions. Among those reliable regions, right fusiform, right rostral ACC and left superior parietal also showed high familiality as well. Overall, regions with greater magnitude of task-related activations showed higher reliability, and reliabilities were greater for beta weights extracted from significant grayordinates within a parcel, compared to the reliabilities for beta weights across the whole parcel. However, some strongly activated task-relevant regions showed only modest reliabilities (parcels overlapping with parts of the ACC, lateral orbitofrontal, superior frontal, and rostral middle frontal regions). In-scanner movement had a moderate negative effect on reliability.

4.1. TRR of behavioral measures

Behavioral measures showed fair to good test-retest reliabilities (ranging from 0.40 to 0.73). Explosion-related outcomes showed lower TRRs, which may be explained in part by the limited number of balloon explosion trials compared to cashout trials. It is important to note that cashout trials may be better suited to capture trait-like risk attitudes. Explosions occur probabilistically; therefore the total number of consecutive risky decisions (number of pumps) available to the subjects in trials ending with an explosion is censored and may not be fully representative of risk-taking behavior. Although reliable behavioral measures of risk taking can be used as markers of individual differences, investigation of neural correlates is important for understanding the spatial localization and mechanisms of individual differences in decision making under risk. Since maladaptive decision-making and heightened risk-taking propensity is observed across a range of psychiatric disorders including addiction, a better knowledge of the underlying biobehavioral mechanisms and identification of reliable and heritable individual differences can inform prevention efforts and potentially help to identify novel medication targets.

4.2. TRR and familiality of task-related brain activations

TRRs of brain activation magnitude were affected by the analytical approach to the modeling of the BOLD response (categorical vs. parametric). Significant task related activations identified using the categorical design (that looked at average of the event) showed low to good reliabilities across cortical parcels (ICCs ranging from zero to 0.72), whereas most of the activations yielded by the parametric design (that looked at parametric modulation of that event with explosion

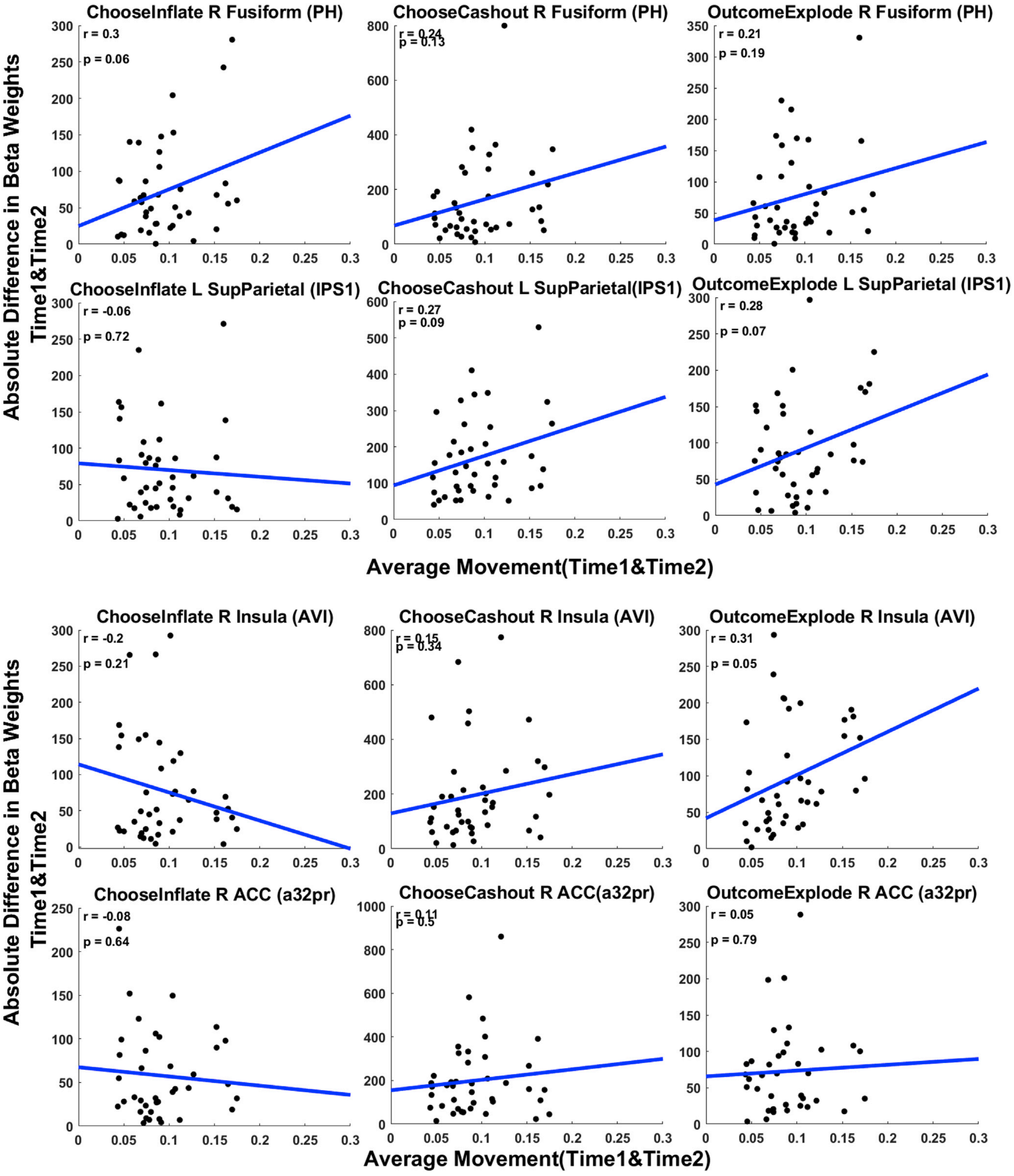


Fig. 7. In-scanner movement does not predict intra-individual variability of activation for ROIs with 'moderate' (top row) and 'low' (bottom row) ICC across sessions. The horizontal axis shows average amount of motion, the vertical axis shows within-subject instability of beta weights (absolute difference between Time1 and Time2) for right Fusiform (PH parcel) and left Superior Parietal regions (IPS1 parcel) (ICCs ranging from 0.36 to 0.59) and for right Insula (AVI parcel) and right ACC regions (a32pr parcel) (ICCs ranging from 0.02 to 0.36) in the *ChooseInflate*, *ChooseCashout* and *OutcomeExplode* contrasts of the categorical design. Note that activation in these regions did not survive whole-brain grayordinate-wise FDR correction in the *OutcomeInflate* contrast only. Each data-point represents a participant.

probabilities) had non-significant reliabilities. Therefore, for the task-related activations, we focus our discussion on TRR findings in the categorical design contrasts (see below for a discussion of possible explanations for the lower TRRs of the parametric design).

In the categorical design, regions that were of *a priori* theoretical importance due to their reported associations with decision making (i.e., insula, OFC, ACC, DLPFC, and caudate) showed significant task-related activations but poor reliabilities. Specifically, only 21 out of 106 selected regions across all four contrasts in the categorical design showed statistically significant reliabilities (when correcting for multiple comparisons). Except for higher reliability observed for the parcel that overlaps with parts of the left anterior insula (AAIC, TRR ICC = 0.54), other parcels that intersect with the insula (left and right FOP4, left and right AVI) had reliabilities ranging from none to poor (0.03–0.32) during risky decisions (*ChooseInflate*). The insula has been reported as one of the most prominent region involved in risky decision making. The same region was also active during risk-averse decisions (*ChooseCashout*, right MI and right FOP4) and had poor reliabilities (ranging from 0.22 to 0.40). One study that looked at ICCs in three ROIs during a Monetary Incentive delay task (n = 13) reported ICCs ranging from 0.47 to 0.63 across different task conditions, but only for the right anterior insula activity (Wu et al., 2014). Parcels that intersect with ACC (a24pr, a24, p24) were active during risky and risk-averse decisions and during evaluation of negative outcomes (*OutcomeExplode*) with reliabilities ranging from poor-to-fair (0.20–0.47). Previous research looking at the ICC in *a priori* ROI of ACC region during an Emotional Faces Interference Task (n = 23) reported poor stability under angry distractors conditions (ranging from −0.005 to 0.022 across high and low perceptual load) but moderate stability under fearful distractors in the low perceptual load condition (ICC 0.54), but not under high perceptual load (Bunford et al., 2017). Parcels that intersect with the DLPFC region (left and right IFSa) were active only during risk-averse decisions and evaluation of negative outcomes and had poor-to-good reliabilities (ranging from 0.34 to 0.56). Other parcels that intersect with DLPFC (9_46d and 46) were active in all contrasts and had poor-to-good reliabilities (parcel 46 TRR ICCs = 0.12 to 0.49, parcel 9_46d TRR ICCs = 0.08 to 0.41). Similarly, Qu et al. (2015a) showed poor reliability of an *a priori* selected the DLPFC region in adolescents (n = 23) when receiving rewards (ICC = 0.34). Besides the regions with *a priori* importance listed above, superior frontal (during risky decisions), posterior cingulate, rostral middle frontal, parstriangularis (during risk averse decisions), rostral/caudal middle frontal (during positive outcome evaluation), and superior/rostral middle frontal (during negative outcome evaluation) had the highest test-retest reliabilities (all TRR ICCs >0.5).

Among all regions that showed significant task-related activations, the fusiform, superior and inferior parietal, and lateral occipital areas had the largest TRRs overall (ranging from moderate-to-high) and were among the ones with high familiality during risky and risk-averse decisions. Interestingly, activations in these regions were not contrast specific and were seen in multiple contrasts. Thus, they may represent non-task specific regions of activation consistently observed in most sensory motor tasks requiring visual attention. Regions with high reliability and familial effects were mostly identified during risky (*ChooseInflate*) and risk-averse decisions (*ChooseCashout*) with the use of the categorical design. These regions included right fusiform, right rostral ACC and left superior parietal regions. More interestingly, although most of the parcels with significant activations had greater TRRs (and also parcels with the greatest TRRs often showed significant task related activations), only a fraction of these parcels with task-related significant activations were among those selected as region of interests based on the prior literature.

The majority of previous decision making studies selected their regions based on theoretical importance (i.e., ACC involvement in conflict processing), rather than their reliabilities. However our findings suggests that majority of literature based ROIs have less than ideal reliabilities. Therefore, studies restricting their analysis to a few regions based on published studies of the task of interest may be basing their conclusions on regions with inherently low reliabilities. In contrast, the regions that are proposed in the current manuscript are based on their reliabilities. With that in mind, although we estimated reliabilities for the regions utilized in previous literature, we did not limit our analysis to a set of theoretical ROIs, but also provided reliabilities for the whole brain at parcel level (see Section 4.3).

It is relevant to note that most of the BART studies are conducted with small (~10) to medium (~70) sample sizes. Individual differences studies that utilize the regions identified in studies with small samples as brain-based biomarkers in studies heritability, development, neurodegeneration or treatment outcomes may result in the use of regions that have unstable of inhomogeneous activity patterns. In addition to reliability studies such as ours, ongoing big-data initiatives in neuroimaging field (such as the Adolescent Brain Cognitive Development and the Human Connectome Project) may contribute to the identification of regions with robust activation patterns and higher reliabilities.

In regard to the BART task, our results indicate that neural activations during the decision period of risk taking in general had greater reliabilities and might be more suitable for addressing the aforementioned research questions, especially when the analysis is not limited to a handful of ROIs. Traditional approaches of ROI selection procedures can go beyond the application of region selection solely based on the theoretical importance of the region and can take into account other factors that affect reliabilities, as discussed in the following sections where we pivot to discussing the results from the *unthresholded parcels* analyses.

4.3. TRR of whole-brain parcellation and its relationship to familiality

Earlier research focusing on ICCs of task fMRI measures revealed a wide range of reliability estimates (mean ICCs ranging from −0.16 to 0.88), with the mean ICC across tasks being 0.5 (Bennett and Miller, 2010). It is quite possible that this mean is biased upwards by a selection bias to report ROIs with higher ICCs among the reviewed studies. While consensus for evaluating ICCs in imaging data does not exist, it seems that values of 0.4–0.5 (or higher) are generally considered an “acceptable” level of reliability for fMRI measures. It is important to note that reliability estimates in most studies are based on a selected number of ROIs, shorter test-retest intervals and a smaller number of subjects compared to the current investigation, in which the whole-brain parcel level reliability averaged about 0.2.

Our study revealed that choice related contrasts had greater number of parcels with moderate reliabilities compared to outcome related contrasts. Within two choice related contrasts, majority of the same parcels showed moderate reliabilities, ones with the highest reliabilities including superior parietal, lateral occipital regions, fusiform, posterior cingulate areas. One possible explanation for greater reliability in the choice related contrasts might be due to a greater number of trials modeled in the fMRI. To investigate that possibility we examined the correlations between the number of trials included in each contrast and the stability of beta weights (i.e. absolute difference in betas between Time1 and Time2) from the *unthresholded parcels* analysis in each corresponding contrast (see *Supplementary Materials* Figure S16). We expected increased stability as the number of events increased. This was indeed the case for the *OutcomeExplode* contrast in both the categorical and parametric designs. However, a statistically significant relationship in the other direction was observed for the *OutcomeInflate* and both of the choice contrasts in the parametric design, which was unexpected and not easily explained. Overall, the relationship between number of events and beta stability seemed variable and contrast dependent. The contrast specificity of the observed moderately reliable regions might be due to how activation changes over sessions for different processes and is an interesting subject to study in future studies. Moreover, the only subcortical region with reliability that passed significance thresholding was the right caudate during risk averse decisions (*ChooseCashout* contrast).

At the whole brain level (across all parcels/segments), the overall degree of reliability was correlated with the overall degree of familiality. However, not all parcels that had high reliability also had high familiality. In the categorical design, cortical regions showing higher reliabilities also showed stronger familial effects for risky and risk-averse decisions; and subcortical regions showing higher reliability also showed stronger familial effects for the risk-averse decisions.

4.4. Factors affecting test-retest reliabilities of neural activations

The first factor that was investigated was the comparison of the reliability estimates of activations from the categorical vs parametric designs. Regardless of how the TRRs were estimated (at parcel level or for specific activated regions), brain activations identified with the categorical design tended to have greater test-retest reliabilities compared to the parametric design, and this difference was statistically significant for the *unthresholded parcels* analysis (Since the *thresholded parcels* do not exactly agree across the two designs, a statistical comparison could not be performed for that analysis). Use of the parametric design in the BART task allows detection of brain activation modulated by the explosion probability. However, parametric designs in general might be less sensitive (also known as low contrast to noise ratio, CNR). As can be seen in Figure S18 (*Supplementary Materials*), the CNR values were greater in the categorical compared to the parametric design, in all contrasts. The degree of systematic change across different levels of a factor would be expected to be much smaller than the difference between the average activity for that factor versus baseline. One factor that might play a role is that the relationship between experimental parameters and the hemodynamic response may have a non-linear relationship or may vary region-to-region (Buchel et al., 1998). Therefore, although the use of parametric designs allows us to investigate the modulated intensity of the cognitive process (Amaro and Barker, 2006), researchers that aim to compare neural activity across multiple sessions with the use of parametric designs may need to try to increase poor reliabilities of the neural activations by improving the sensitivity of parametric designs with approaches such as alternative modeling of hemodynamic response (i.e., nonlinear regressors), collecting more data, or other approaches.

The second factor that we investigated was the level of activation at first scan. The magnitude and the effect size of regional activations at first scan were correlated with TRRs of cortical parcels in all conditions except the evaluation period of positive outcomes (*OutcomeInflate* contrast) in the categorical design and risky decisions (*ChooseInflate* contrast) in the parametric design. The magnitude and effect size were not correlated with reliabilities in subcortical regions, perhaps due to the lower tSNR in subcortical regions. These findings are in line with the findings of Caceres et al. (2009) showing a relationship between activation level and reliabilities, albeit in a region dependent manner. Moreover, Raemaekers et al. (2007) investigated the mechanisms of how temporal signal-to-noise ratio (tSNR) can influence reliability estimates and demonstrated that between subject variation in brain activation can be explained to some degree by between subject variations in tSNRs (r ranging from 0.73 to 0.91, dependent on contrast), with intra-subject tSNRs highly reliable. Analogously, the reliability differences across parcels observed in the current study showed moderate positive correlations with the tSNR differences across these parcels. More interestingly, after regressing average motion from the parcel level tSNRs, the relationship between ICCs and tSNR disappeared (see Supplementary Figure S15), suggesting that this relationship was in part driven by differences in the amount of motion. With increasing levels of motion, the tSNR decreased and square root of the variance across timepoints increased across the whole brain (see Supplementary Materials Figure S19), which in return resulted in lower reliabilities.

Our finding that regions with greater activation had greater reliabilities complements our finding that parcels containing only active grayordinates had greater reliabilities compared to the reliabilities estimated for the whole parcel. Earlier research has shown that ICC values are greater for larger parcels/segments compared to smaller ones (Shah et al., 2016). This may be driven by the fact that larger parcels/segments have a greater number of data points for averaging, possibly increasing signal-to-noise ratio in the averaged measure. In our study, activated clusters were further divided into anatomical parcels/segments, therefore all of our *thresholded parcels* were smaller in size than the *unthresholded parcels*. However, despite this size difference, our results showed that *thresholded parcels* had greater ICCs compared to *unthresholded parcels*. Therefore, the choice of method for defining an ROI (by using parcel boundaries only versus limiting the ROI to the significant activity within the parcel) should be evaluated on a case by case basis (i.e., the spatial extent of the ROI, tSNR in the parcel, type of parcellation - structural vs functional). Overall, our findings indicate that the magnitude of activation is at least one of the factors that contribute to TRR.

Lastly, we investigated the effect of motion on consistency of activation magnitude. Overall, motion had a moderate negative affect on the consistency of beta values across sessions, therefore studies in which greater subject movement is anticipated (e.g., studies including children, or individuals with ADHD) might aim to implement more advance motion detection and correction algorithms. It is important to note that the effect of motion on the consistency of brain activations, while broadly evident across the whole brain, was weaker for individual parcels, regardless of whether the individual parcels had high or low reliabilities.

4.5. Limitations

We acknowledge that given the lack of dizygotic twin pairs in our study, we cannot distinguish between genetic and environmental influences in our estimate of ‘familiality’. Another concern is that the dependencies in the data introduced due to the MZ twins could potentially bias the ICC estimates, since the ICC model did not concurrently model the sibships. In order to investigate this possibility, we re-estimated reliabilities by assigning Twin 1 and Twin 2 of our twin pairs to separate groups, which resulted in two samples with no dependencies (unrelated individuals). The ICC values averaged over the two independent samples were very similar to those derived using the full sample, with a regression line nearly indistinguishable from the line of identity, indicating that there was no evidence that the MZ twins biased the ICC estimates in any systematic fashion (Supplementary Figure S20). Lastly, since ICCs are always inherent to a specific sample, we acknowledge that our TRR estimates will be dependent on our study cohort, study design (event-related), duration between sessions, and scanning parameters and may not generalize to other populations, studies utilizing block designs or an entirely different scanning or analysis protocol. Nonetheless, our results importantly inform the expected reliability of the BART specifically, and other fMRI decision/risk paradigms more generally.

5. Conclusions

Maladaptive decision making has been implicated in many psychiatric disorders, including substance use disorder, depression, bipolar disorder and schizophrenia, but is also related to poor real-life outcomes in healthy individuals (Caceda et al., 2014). The reliable regions identified in this study are good candidate for use in clinical neuroscience research. The range of reliabilities of fMRI measures varies greatly across constructs. Therefore, reliabilities of fMRI measures reported in this study are specific to the BART task. Future individual differences studies are advised to choose a paradigm for which reliable neural markers have been identified in adequately powered test-retest reliability studies. With the recent trend toward collaborative, large-sample studies aimed at the generation of “big data”, this step can help to ensure that the resulting data are adequate for addressing research questions that assume the existence of reliable inter-individual differences in brain activation. For a novel paradigm, an estimation of test-retest reliability is recommended before a paradigm is used in a genetic, clinical, or developmental study.

However we also acknowledge that it is not always feasible for every fMRI study to estimate their own reliabilities. Therefore we now provide general recommendations based on the BART study. Based on the factors investigated in this study, we recommend that decision making paradigms utilize the activations identified in the contrast that taps into the decision rather than outcome related processes for use as brain-based biomarkers in studying heritability, development, neurodegeneration and treatment outcomes. We base this recommendation on the fact that there were considerably more regions with good reliability in the choice related contrasts compared to the outcome related contrasts. Notably, the only subcortical region with a significant reliability was the right caudate (in the *ChooseCashout* contrast). The generally poor reliabilities of the subcortical regions in our study is possibly a consequence of the low temporal SNR subcortically (Figure S17), which itself is primarily a consequence of the acquisition voxel size. Moreover, individual difference studies using fMRI paradigms for which the reliability is not explicitly known should take into account reliability estimates reported in previous studies as part of their ROI selection (rather than just theoretical relevance), and can also benefit from estimating tSNRs and selecting regions with the highest activation levels. We recommend that the reliability of task-related brain activations, particularly in parametric designs, be firmly established before they are employed for studies concerned with individual differences such as investigation of correlations between brain activation and behavioral or clinical variables, or longitudinal studies.

In assessment of TRR itself, one special concern might be in regard to developmental studies and studies focusing on cognitive decline, in which fMRI measures might appear less stable due to nonlinear changes in decision making or differences in an individual's position along developmental or degenerative trajectories, which becomes more of an issue the farther the test and retest sessions are separated in time. To overcome that, developmental studies can implement a careful recruitment protocol, matching for age or pubertal status via a developmental biomarker. Studies focusing on cognitive decline in aging might benefit from considering the neurodegenerative trajectory of their subjects.

Declaration of competing interest

All authors declare that they have no conflict of interest.

CRediT authorship contribution statement

Ozlem Korucuoglu: Data curation, Formal analysis, Writing - original draft, Methodology. **Michael P. Harms:** Methodology, Writing - review & editing. **Serguei V. Astafiev:** Data curation, Methodology, Formal analysis, Writing - review & editing. **James T. Kennedy:** Data curation, Formal analysis, Writing - review & editing. **Semyon Golosheykin:** Data curation, Formal analysis. **Deanna M. Barch:** Methodology, Writing - review & editing. **Andrey P. Anokhin:** Conceptualization, Funding acquisition, Supervision, Writing - review & editing.

Acknowledgements

Research reported in this publication was supported by the National Institute on Drug Abuse (NIDA) and by Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) of the National Institutes of Health under award numbers R21DA038834, R01DA018899, and R01HD083614 to A.P.A. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors acknowledge organizational and technical support the project staff. The authors also acknowledge the generous giving of time by the study participants.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2020.116759.

References

Amaro, E., Barker, G.J., 2006. Study design in fMRI: basic principles. Brain Cognit. 60, 220–232. https://doi.org/10.1016/j.bandc.2005.11.009.

Bennett, C.M., Miller, M.B., 2013. fMRI reliability: influences of task and experimental design. Cognit. Affect Behav. Neurosci. 13, 690–702. https://doi.org/10.3758/s13415-013-0195-1.

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci. 1191, 133–155. https://doi.org/10.1111/j.1749-6632.2010.05446.x.

Buchel, C., Holmes, A.P., Rees, G., Friston, K.J., 1998. Characterizing stimulus – response functions using nonlinear regressors in parametric fMRI experiments. Neuroimage 8, 140–148.

Bunford, N., Kinney, K.L., Michael, J., Klumpp, H., 2017. Threat distractor and perceptual load modulate test-retest reliability of anterior cingulate cortex response. Prog. Neuro-Psychopharmacol. Biol. Psychiatry 77, 120–127.

Caceda, R., Nemeroff, C., Harvey, P.D., 2014. Toward an understanding of decision making in severe mental illness. J. Neuropsychiatry Clin. Neurosci. 26, 196–213.

Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C.R., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage 45, 758–768. https://doi.org/10.1016/j.neuroimage.2008.12.035.

Casey, B.J., Cannonier, T., Conley, M.I., Cohen, A.O., Barch, D.M., Heitzeg, M.M., Soules, M.E., Teslovich, T., Dellarco, D.V., Garavan, H., Orr, C.A., Wager, T.D., Banich, M.T., Speer, N.K., Sutherland, M.T., Riedel, M.C., Dick, A.S., Bjork, J.M., Thomas, K.M., Chaarani, B., Mejia, M.H., Hagler, D.J., Cornejo, M.D., Sicat, C.S., Harms, M.P., Dosenbach, N.U.F., Rosenberg, M., Earl, E., Bartsch, H., Watts, R., Polimeni, J.R., Kuperman, J.M., Fair, D.A., Dale, A.M., 2018. The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. Dev. Cogn. Neurosci. 32, 43–54. https://doi.org/10.1016/j.dcn.2018.03.001.

Chung, S., You, I., Cho, G., Chung, G., Shin, Y., Kim, D., Choi, S., 2009. Changes of functional MRI findings in a patient whose pathological gambling improved with fluvoxamine. Yonsei Med. J. 50, 441–444. https://doi.org/10.3349/ymj.2009.50.3.441.

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

Coalson, T.S., Essen, D.C. Van, Glasser, M.F., 2018. The impact of traditional neuroimaging methods on the spatial localization of cortical areas. Proc. Natl. Acad. Sci. Unit. States Am. 115, E6356–E6365. https://doi.org/10.1073/pnas.1801582115.

Elliott, M.L., Knodt, A.R., Ireland, D., Morris, M.L., Poulton, R., Ramrakha, S., Sison, M.L., Moffitt, T.E., Caspi, A., Hatriri, A.R., 2019. Poor test-retest reliability of task-fMRI: new empirical evidence and a meta-analysis. BioRxiv. https://doi.org/10.1101/681700.

Engelhardt, L.E., Abbe, M., Juranek, J., Demaster, D., Harden, K.P., Tucker-drob, E.M., Church, J.A., 2017. Children's head motion during fMRI tasks is heritable and stable over time. Dev. Cogn. Neurosci. 25, 58–68. https://doi.org/10.1016/j.dcn.2017.01.011.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Kouwe, A. Van Der, Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: neurotechnique automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.

Frohner, J.H., Teckentrup, V., Smolka, M.N., Kroemer, N.B., 2019. Addressing the reliability fallacy in fMRI: similar group effects may arise from unreliable individual effects. Neuroimage 195, 174–189.

Fukunaga, R., Brown, J.W., Bogg, T., 2012. Decision making in the Balloon Analogue Risk Task (BART): anterior cingulate cortex signal loss-aversion but not the infrequency of risky choices. Cognit. Affect Behav. Neurosci. 12 (3), 479–490.

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., Smith, S.M., Van Essen, D.C., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178. https://doi.org/10.1038/nature18933.

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., Essen, D.C. Van, Jenkinson, M., Hcp, W., 2013. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage 80, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127.

Gorgolewski, K.J., Storkey, A., Bastin, M.E., Whittle, I.R., Wardlaw, J.M., Pernet, C.R., 2013a. A test-retest fMRI dataset for motor, language and spatial attention functions. GIGA Sci. 2, 2–5.

Gorgolewski, K.J., Storkey, A.J., Bastin, M.E., Whittle, I., Pernet, C., 2013b. Single subject fMRI test–retest reliability metrics and confounding factors. Neuroimage 69, 231–243. https://doi.org/10.1016/j.neuroimage.2012.10.085.

Hulvershorn, L.A., Hummer, T.A., Fukunaga, R., Leibenluft, E., Finn, P., Cyders, M.A., et al., 2015. Neural activation during risky decision-making in youth at high risk for substance use disorders. Psychiatry Res. Neuroimaging. 233 (2), 102–111.

Insel, C., Somerville, L.H., 2018. Asymmetric neural tracking of gain and loss magnitude during adolescence. Soc. Cognit. Affect Neurosci. 13, 785–796. https://doi.org/10.1093/scan/nsy058.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. FSL. Neuroimage. 62, 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015.

Kendler, K., Neale, M., 2009. "Familiality" or heritability. Arch. Gen. psychiat. 66, 452–453. https://doi.org/10.1001/archgenpsychiatry.2009.14.

Kohno, M., Ghahremani, D.G., Morales, A.M., Robertson, C.L., Ishibashi, K., Morgan, A.T., et al., 2015. Risk-taking behavior: dopamine D2/D3 receptors, feedback, and frontolimbic activity. Cerebr. Cortex 25 (1), 236–245. https://doi.org/10.1093/cercor/bht218.

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

Korucuoglu, O., Harms, M.P., Kennedy, J.T., Golosheykin, S., Astafiev, S.V., Barch, D.M., Anokhin, A.P., 2019. Adolescent decision-making under risk: Neural correlates and sex differences. Cereb Cortex. https://doi.org/10.1093/cercor/bhz269.

Krain, A.L., Wilson, A.M., Arbuckle, R., Castellanos, F.X., Milham, M.P., 2006. Distinct neural mechanisms of risk and ambiguity: a meta-analysis of decision-making. Neuroimage 32, 477–484. https://doi.org/10.1016/j.neuroimage.2006.02.047.

Macoveanu, J., Fisher, P.M., Haahr, M.E., Frokjaer, V.G., Knudsen, G.M., Siebner, H.R., 2014. Effects of selective serotonin reuptake inhibition on neural activity related to risky decisions and monetary rewards in healthy males. Neuroimage 99, 434–442. https://doi.org/10.1016/j.neuroimage.2014.05.040.

Mccrae, R.R., Kurtz, J., Yamagata, S., Terracciano, A., 2011. Internal consistency, retest reliability, and their implications for personality scale validity. Pers. Soc. Psychol. Rev. 15, 28–50. https://doi.org/10.1177/1088868310366253.Internal.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlations coefficients. Psychol. Methods 1, 30–46. https://doi.org/10.1037/1082-989X.1.4.390.

Qu, Y., Fuligni, A.J., Galvan, A., Telzer, E.H., 2015a. Buffering effect of positive parent-child relationships on adolescent risk taking: a longitudinal neuroimaging investigation. Dev. Cog. Neurosci. 15, 26–34. https://doi.org/10.1016/j.dcn.2015.08.005.

Qu, Y., Galvan, A., Fuligni, A.J., Lieberman, M.D., Telzer, E.H., 2015b. Longitudinal changes in prefrontal cortex activation underlie declines in adolescent risk taking. J. Neurosci. 35, 11308–11314. https://doi.org/10.1523/JNEUROSCI.1553-15.2015.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL. https://www.R-project.org/.

Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J.A., Kahn, R.S., Ramsey, N.F., 2007. Test-retest reliability of fMRI activation during prosaccades and antisaccades. Neuroimage 36, 532–542. https://doi.org/10.1016/j.neuroimage.2007.03.061.

Rao, H., Korczykowski, M., Pluta, J., Hoang, A., Detre, J., 2008. Neural correlates of voluntary and involuntary risk taking in the human brain: an fMRI Study of the Balloon Analog Risk Task (BART). Neuroimage 42 (2), 902–910. https://doi.org/10.1016/j.neuroimage.2008.05.046.

Rao, L., Zhou, Y., Zheng, D., Yang, L., 2018. Genetic contribution to variation in risk taking : a functional MRI twin study of the Balloon Analogue Risk Task. Psychol. Sci. 29, 1679–1691. https://doi.org/10.1177/0956797618779961.

Robinson, E.C., Garcia, K., Glasser, M.F., Chen, Z., Coalson, T.S., Makropoulos, A., Bozek, J., Wright, R., Schuh, A., Webster, M., Hutter, J., Price, A., Grande, L.C., Hughes, E., Tusor, N., Bayly, P.V., Van Essen, D.C., Smith, S.M., Edwards, A.D., Hajnal, J., Jenkinson, M., Glocker, B., Rueckert, D., 2018. Multimodal surface matching with higher-order smoothness constraints. Neuroimage 167, 453–465. https://doi.org/10.1016/j.neuroimage.2017.10.037.

Rock, P L, Roiser, J P, Riedel, W J, Blackwell, A D, 2014. Cognitive impairment in depression: a systematic review and meta-analysis. Psychol. Medicine 44, 2029–2040. https://doi.org/10.1017/S0033291713002535.

Shah, L.M., Cramer, J.A., Ferguson, M.A., Birn, R.M., Anderson, J.S., 2016. Reliability and reproducibility of individual differences in functional connectivity acquired during task and resting state. Brain Behav. 6, e00456 https://doi.org/10.1002/brb3.456.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428.

Silverman, M.H., Jedd, K., Luciana, M., 2015. Neural networks involved in adolescent reward processing: an activation likelihood estimation meta-analysis of functional neuroimaging studies. Neuroimage 122, 427–439. https://doi.org/10.1016/j.neuroimage.2015.07.083.

Termenon, M., Jaillard, A., Delon-Martin, C., Achard, S., 2016. Reliability of graph analysis of resting state fMRI using test-retest dataset from the Human Connectome Project. Neuroimage 142, 172–187. https://doi.org/10.1016/j.neuroimage.2016.05.062.

Tisdall, M.D., Hess, A.T., Reuter, M., Meintjes, E.M., Fischl, B., van der Kouwe, A.J.W., 2012. Volumetric navigators for prospective motion correction and selective reacquisition in neuroanatomical MRI. Magn. Reson. Med. 68, 389–399. https://doi.org/10.1002/mrm.23228.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. Neuroimage 92, 381–397.

Winkler, A.M., Webster, M.A., Vidaurre, D., Nichols, T.E., Smith, S.M., 2015. Multi-level block permutation. Neuroimage 123, 253–268.

Wu, C.C., Smanez-Larkin, G.R., Katovich, K., Knutson, B., 2014. Affective traits link to reliable neural markes of incentive anticipation. Neuroimage 84, 279–289.

Zeng, L., Wang, D., Fox, M.D., Sabuncu, M., Hu, D., Ge, M., Buckner, R.L., Liu, H., 2014. Neurobiological basis of head motion in brain imaging. Proc. Natl. Acad. Sci. Unit. States Am. 111, 6058–6062. https://doi.org/10.1073/pnas.1317424111.