Published in final edited form as:
Neuroimage. 2020 April 01; 209: 116495. doi:10.1016/j.neuroimage.2019.116495.

Test-retest reliability of brain responses to risk-taking during the balloon analogue risk task

Xiong Li[a,1], **Yu Pan**[a,b,1], **Zhuo Fang**[b,c], **Hui Lei**[c], **Xiaocui Zhang**[c], **Hui Shi**[c], **Ning Ma**[c], **Philip Raine**[c], **Reagan Wetherill**[d], **Junghoon J. Kim**[e], **Yan Wan**[a], **Hengyi Rao**[b,c,*]

[a]School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing, China

[b]Key Laboratory of Applied Brain and Cognitive Sciences, School of Business and Management, Shanghai International Studies University, Shanghai, China

[c]Center for Functional Neuroimaging, Department of Neurology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

[d]Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

[e]Department of Molecular, Cellular, and Biomedical Sciences, CUNY School of Medicine, The City College of New York, New York, NY, USA

Abstract

The Balloon Analogue Risk Task (BART) provides a reliable and ecologically valid model for the assessment of individual risk-taking propensity and is frequently used in neuroimaging and developmental research. Although the test-retest reliability of risk-taking behavior during the BART is well established, the reliability of brain activation patterns in response to risk-taking during the BART remains elusive. In this study, we used functional magnetic resonance imaging (fMRI) and evaluated the test-retest reliability of brain responses in 34 healthy adults during a modified BART by calculating the intraclass correlation coefficients (ICC) and Dice's similarity coefficients (DSC). Analyses revealed that risk-induced brain activation patterns showed good test-retest reliability (median ICC = 0.62) and moderate to high spatial consistency, while brain activation patterns associated with win or loss outcomes only had poor to fair reliability (median

[*]Corresponding author. Center for Functional Neuroimaging & Department of Neurology, University of Pennsylvania Perelman School of Medicine, Room D502, Richards Medical Research Building, 3700 Hamilton Walk, Philadelphia, PA, 19104, USA. hengyi@pennmedicine.upenn.edu (H. Rao).

[1]Li and Pan contributed equally to this work.

CRediT authorship contribution statement

Xiong Li: Writing - original draft, Methodology, Formal analysis. **Yu Pan:** Conceptualization, Supervision, Writing - original draft, Funding acquisition. **Zhuo Fang:** Data curation, Formal analysis, Writing - review & editing. **Hui Lei:** Formal analysis, Methodology, Writing - review & editing. **Xiaocui Zhang:** Visualization, Writing - review & editing. **Hui Shi:** Data curation, Project administration, Writing - review & editing. **Ning Ma:** Data curation, Project administration. **Philip Raine:** Writing - review & editing. **Reagan Wetherill:** Writing - review & editing. **Junghoon J. Kim:** Writing - review & editing. **Yan Wan:** Writing - review & editing. **Hengyi Rao:** Conceptualization, Supervision, Project administration, Writing - review & editing, Funding acquisition.

Appendix A. Supplementary data
Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.116495.

ICC = 0.33 for win and 0.42 for loss). These findings have important implications for future utility of the BART in fMRI to examine brain responses to risk-taking and decision-making.

Keywords

Test-retest reliability; Balloon analogue risk task (BART); fMRI; Intraclass correlation coefficient (ICC)

1. Introduction

The Balloon Analogue Risk Task (BART), a computerized and laboratory-based paradigm originally developed by Lejuez et al. (2002), is widely used for measuring individual risk-taking behavior in various studies. During the BART, risk-taking propensity is defined as the average adjusted pump (i.e. average number of balloon inflation pumps in the win trials), which is directly linked to the probability of explosion for each balloon, and the outcome may influence and modulate the individual's next decision (Lejuez et al., 2002; Xu et al., 2016, 2018). As such, the BART parallels real-world risky behaviors (Lauriola et al., 2014). Given the advantage of high ecological validity, numerous behavioral studies have used the BART to investigate individuals' risk-taking performance and to examine associations between BART performance and real-world behavior (Aklin et al., 2005; Hopko et al., 2006; Kathleen Holmes et al., 2009; Lauriola et al., 2014; Lejuez et al., 2002, 2003; 2005, 2007; MacPherson et al., 2010; MacLean et al., 2018).

In addition to behavioral studies, the use of BART in neuroimaging studies has increased significantly over the past decade. For example, Rao et al. (2008) conducted the first BART neuroimaging study in which they modified Lejuez et al. (2002) to be compatible with the scanner environment and found that risk processing during the BART was associated with robust activation in the mesolimbic-frontal reward system, including the thalamus/midbrain, ventral and dorsal striatum, anterior insula, dorsal lateral prefrontal cortex (DLPFC), and anterior cingulate/medial frontal cortex (ACC/MFC). Since then, the BART has been increasingly employed in neuroimaging research to study the neural correlates of risk-taking and decision-making. Previous studies have successfully used this task to: (1) explore the relationship between BART-induced neural activation and behaviors or traits (e.g., Claus and Hutchison, 2012; Galván et al., 2013; Helfinstein et al., 2014; Pan et al., 2018; Qu et al., 2015; Xu et al., 2016, 2018); (2) investigate the influences of altered psycho-physiological processes such as medication and sleep deprivation on BART-related brain activation change (e.g., Lee et al., 2009; Lei et al., 2017); and (3) compare brain activation differences between different cohorts or clinical populations during risk-taking and decision-making (Fang et al., 2017; Lighthall et al., 2009; Qi et al., 2015; Rao et al., 2011).

One advantage of the BART paradigm is that the behavioral performance may serve as a trait-like characteristic to study individual differences in personality (Bornovalova et al., 2009; Ko et al., 2010; Mishra and Novakowski, 2016; Parkinson et al., 2012) and real-world risk-taking behaviors (Aklin et al., 2005; Lejuez et al., 2003; MacPherson et al., 2010). Moreover, the BART can also be administered repeatedly with minimal practice effects, making it an ideal paradigm for longitudinal and developmental research (Banducci et al., 2015; Reynolds et al., 2014; Braams et al., 2015). For example, a recent longitudinal BART study conducted annually over three years found that maternal risk-taking propensity could predict youth's alcohol use (Banducci et al., 2015). Another study demonstrated significant peer influence on risk-taking behavior in adolescents (Reynolds et al., 2014). Furthermore, Braams et al. (2015) used the BART to examine longitudinal changes in adolescent risk-taking and found a quadratic age pattern for nucleus accumbens activity in response to rewards. However, future utility of the BART in longitudinal and developmental research is predicated on the assumption of good test-retest reliability of the BART behavior and its associated brain responses.

Previous behavioral studies have consistently demonstrated good test-retest reliability of BART performance. For example, Lejuez et al. (2003) examined the reliability of the BART among 26 smokers and 34 nonsmokers with an age range of 18–30 years. By administering the BART three times on a single day, the authors found correlations between BART performances ranging from 0.62 to 0.82. White et al. (2008) assessed the reliability of BART performance in 39 adults with an age range of 18–35 years across three sessions and reported moderate to high correlations (r = 0.66 to 0.78). Xu et al. (2013) compared the reliability of the BART to the delay discounting task (DDT) and the Iowa gambling task (IGT), and reported similar correlation coefficients (r = 0.66 to 0.76) for the BART across three sessions within 2-week intervals in 40 college students with an age range of 19–22 years, which was higher than the ICC of DDT (r = 0.35 to 0.65) and IGT (r = 0.53 to 0.67). In addition, Weafer et al. (2013) investigated the reliability of the BART in a large sample of 119 adults with an age range of 18–30 years and reported a correlation of r = 0.79 across repeated tests with an average interval of 8.6 days. Taken together, the moderate to high test-retest stability of the BART risk-taking behavior indicates that one's BART performance on a single occasion may be representative of the individual's general risk performance. However, the test-retest reliability of brain responses during the BART remains unclear.

Using functional magnetic resonance imaging (fMRI), early studies consistently suggest that imaging measures may be more sensitive to detect changes over time or differences between groups than behavioral measures (Callicott et al., 2003; Raemaekers et al., 2006, 2007; Vink et al., 2006). Similarly, several recent BART imaging studies observed significant brain activation differences but failed to find behavioral differences between comparison groups (Fang et al., 2017; Galván et al., 2013; Rao et al., 2011). Given the high sensitivity of fMRI to detect differences, researchers have started to use brain activation measured by fMRI as a biomarker in relation to healthy aging (Dosenbach et al., 2010; Geerligs et al., 2014), personality (Yarkoni, 2015), intelligence (van den Heuvel et al., 2009), and mood (Smith et al., 2015) (for a review, see Dubois and Adolphs, 2016). However, the conclusions of fMRI studies highly depend on its reproducibility and reliability, which has increasingly become a key concern in the field (e.g., Bennett and Miller, 2010; Chase et al., 2015; Dubois and Adolphs, 2016; Schwartz et al., 2019). If fMRI findings show a high degree of reproducibility and reliability, they can contribute solid scientific knowledge, be generalized, and provide a meaningful interpretation of functional changes across time, especially for developmental research. Reliability is also important for large-scale research where findings across different scanners, populations, and sites are compared. Furthermore, reliability must be established before fMRI can be used for medical or legal applications (Bennett and

Miller, 2010). As such, current literature on the reliability of task-induced brain activation studies (e.g., Atri et al., 2011; Brandt et al., 2013; Cao et al., 2014; Chase et al., 2015; Gorgolewski et al., 2015; Gorgolewski et al., 2013; Morrison et al., 2016; Nettekoven et al., 2018; Plichta et al., 2012; Raemaekers et al., 2012; Sauder et al., 2013; Yang et al., 2019) as well as the reliability of resting-state brain imaging studies (e.g., Blautzik et al., 2013; Braun et al., 2012; Chen et al., 2015; Guo et al., 2012; Li et al., 2012; Liao et al., 2013; Mannfolk et al., 2011; Somandepalli et al., 2015; Song et al., 2012; Yang et al., 2019; Zuo and Xing, 2014) is growing rapidly.

Despite the widespread use of the BART paradigm for assessing risk-taking behavior and brain function, the test-retest reliability of brain responses to the BART has not been evaluated. Therefore, the current study aimed to address this knowledge gap by examining the test-retest reliability of brain responses to risk-taking during the BART. The spatial consistency of brain activation patterns was also assessed. Although not the primary focus of this study, we examined the test-retest reliability of brain response to the processing of loss and win outcomes during the BART. According to previous BART fMRI studies (Claus and Hutchison, 2012; Fang et al., 2017; Galván et al., 2013; Helfinstein et al., 2014; Lei et al., 2017; Pan et al., 2018; Rao et al., 2008), the risk and outcome processing consistently show robust activation in the mesolimbic-frontal pathway, including ACC/MFC, DLPFC, the thalamus/midbrain, ventral and dorsal striatum, and insula, therefore, we expect the activation in crucial hubs of the mesolimbic-frontal pathway will show good test-retest reliability.

2. Methods

2.1. Participants

Participants were recruited in response to study advertisements posted around the university campus. Thirty-four healthy adults (18 females, mean age = 32.5 ± 8.7 yrs) took part in the study and completed the BART during fMRI on two separate occasions. The scan sessions occurred at the same time of the day with an interval of one to three days. Participants were free of major medical conditions and psychiatry disorders, as determined by interviews, clinical history, questionnaires and physical examinations. Participants were non-smokers and were required to provide a negative result on a urine drug test before the study. All participants had at least a high school education. All study procedures adhered to the Declaration of Helsinki and were approved by the University of Pennsylvania Institutional Review Board (IRB). Subjects were provided written, informed, IRB-approved consent prior to participating in the study.

2.2. Balloon analogue risk task

The BART paradigm modified by Rao et al. (2008) was applied in the current study (Fig. 1). During the task, participants were presented a virtual balloon and asked to press a button to inflate the balloon that could either grow larger or explode. As the size of the balloon increased, the associated risk of explosion and the monetary reward increased as well. Subjects had the option to continue or discontinue inflating the balloon by pressing two buttons. If subjects stopped inflating the balloon, they would collect the wager for the current balloon, and the amount of the reward was added to the cumulative earnings. However, if subjects continued to inflate the balloon and the balloon exploded, subjects lost the wager of the current balloon and the amount of the wager was subtracted from the cumulative earnings as the penalty. For each balloon, the maximum number of inflations participants could make was 12, which was unknown to the subjects. The wager size and the probability of explosion both monotonically increased with the number of inflations for each balloon. Specifically, the probability of explosion was set to monotonically increase from 0 to 89.6% and the wager increased from 0 to 5.15 dollars, from the smallest balloon to the largest balloon. The timing of inflation was controlled by a cue, which consisted of a small circle that changed color from red to green with a jittered time interval. Participants could press a button to continue or discontinue inflation only when the color of the cue was green. The cue immediately turned red for a jittered time interval between 1.5 and 2.5s after participants successfully pressed a button and inflated the balloon, and then turned green again to indicate the next inflation. There was also a jittered 2–4 s interval after the end of previous balloon and prior to the beginning of next balloon. The time point when each balloon exploded was manipulated randomly, and participants were unaware of the exact probability of explosion associated with a given inflation. The outcome for each trial was immediately provided to participants once they collected the wager or the balloon exploded. More detailed parameters of the BART can be found in Rao et al. (2008).

2.3. Image data acquisition

MRI scans were conducted in a 3T Siemens Trio scanner (Siemens Medical Systems, Erlangen, Germany). A standard echo-planar imaging (EPI) sequence was used to acquire BOLD fMRI data while participants performed the BART (TR = 1.5 s, TE = 24 ms, FOV = 220 × 220 mm, matrix = 64 × 64 × 25, slices thickness = 4 mm, inter-slice gap = 4 mm, 240 acquisitions). High-resolution anatomical images were obtained using a standard 3D Magnetization Prepared Rapid Acquisition Gradient Echo (MPRAGE) sequence (TR = 1620 ms, time to inversion (TI) = 950 ms, TE = 3 ms, flip angle = 15°, 160 contiguous slices, 1 × 1 × 1 mm resolution).

2.4. MRI data analyses

SPM8 (www.fil.ion.ucl.ac.uk/spm) was used for BART imaging data preprocessing and analyses. Standard preprocessing steps were applied. Functional images were realigned to correct head motion, corrected for slice acquisition time differences, and coregistered with the anatomical image. Functional data were then normalized to the standard MNI brain template with a 2 × 2 × 2 mm^3 voxel size, smoothed with 8 mm FWHM Gaussian kernel, and entered into a voxel-wise analysis using general linear model (GLM). A high-pass filter with a cut-off at 128 s was used to remove low frequency fluctuations. The inclusion criteria of head motion was FD power <0.5 according to Power et al. (2014).

Preprocessed BART image data were modeled using a standard hemodynamic response function (HRF) with time derivative. For the voxel-wise whole brain analysis, the model included three regressors representing three events after participants pressing a button: (1) balloon inflation (i.e. onset of a larger balloon). (2) win outcome (i.e. onset of the win feedback/collected the wager). (3) loss outcome (i.e. onset of the loss feedback/balloon exploded). The probability of explosion for each balloon as the parametric risk level was orthogonalized by mean central correction, and then was entered into the model as a linear parametric modulation of the balloon inflation regressor. Three SPM t-contrast maps were obtained for each participant. (1) The contrast of risk defined to examine brain activation that covaried with the parametric risk level. (2) The contrast of win defined to examine brain activation in response to win outcomes. (3) The contrast of loss defined to examine brain activation in response to loss outcomes.

2.5. Analysis for test-retest reliability of BART behavioral performances

Using Pearson correlations, Spearman rank-order correlations, and intraclass correlation analyses, we examined the test-retest reliability of the BART behavioral performances across the two sessions. SPSS 24.0 (IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp) was used for the analyses (Koo and Li, 2016).

2.6. Analysis for test-retest reliability of brain activation

We first performed group-level analyses using one-sample t-tests. The t-statistics of the three relevant contrasts (i.e. contrast of risk, contrast of win, and contrast of loss) for the two scan sessions were calculated for every voxel. Statistical significance was set at a threshold of uncorrected $p < 0.001$ at the whole-brain level and family-wise error (FWE) corrected $p < 0.05$ at the cluster level for the brain activation associated with increased risk level. Because the number of outcome trials was much less than the number of inflations, the statistical significance was set at a threshold of uncorrected $p < 0.001$ at the whole-brain level for the brain activation in response to win and loss outcomes. The intraclass correlation coefficient (ICC) analysis was then used to measure the test-retest reliability of the individual activation strength of brain response during the BART in 34 subjects across the two fMRI scan sessions. ICC is defined by ratio of the between-subject variance and the total variance (Shrout and Fleiss, 1979), which informs on the ability of fMRI to assess differences in brain activity between subjects (Caceres et al., 2009). Typically, the coefficient ranges from zero (no reliability) to one (perfect reliability), and as such, we used the ICC to assess the stability of inter-individual differences in brain activation magnitude over time (Brandt et al., 2013; McGraw and Wong, 1996). Note that ICCs can be estimated as negative when the true value approaches zero (Murray et al., 1996). For ICCs estimated as negative, researchers can either use the negative ICC or constrain the negative ICC to 0 (Baldwin et al., 2005).

One aim of the current study is to compute the reliability of individual and group-wise neuroimaging results. Based on the ICC definition by Shrout and Fleiss (1979), we applied ICC $(_{3,1})$, which is a within-subject measurement to examine the intraclass correlation of contrast t-value for pairs of activation maps. The mathematical equation of ICC $_{(3,1)}$ in Shrout and Fleiss (1979) was described as:

$$\mathrm{ICC}_{(3,\,1)} = (\mathrm{BMS} - \mathrm{WMS}) / (\mathrm{BMS} + (k - 1)\,\mathrm{WMS})$$

BMS: between-subject variance; WMS: within-subject variance; k: number of raters/ measurements.

The within-subject ICC implementation has been employed by several neuroimaging studies (Brandt et al., 2013; Caceres et al., 2009; Raemaekers et al., 2007; Specht et al., 2003) to assess the test-retest signal across region of interest (ROI) voxels for each subject. This approach measures ICC by computing the amount of total variance explained by intra-voxel variance and test the consistency of the spatial distribution of the BOLD signal in a given region for each individual. The ICC toolbox (https://www.kcl.ac.uk/ioppn/depts/neuroimaging/research/imaginganalysis/Software/ICC-Toolbox.aspx) was used to assess the test-retest reliabilities in specific ROIs and the whole-brain (Caceres et al., 2009). Additionally, the toolbox can produce all three types of ICC maps for any number of sessions. Mean squares maps are produced and ICC values are converted to a Fisher's Z map. The ICC maps can be examined for the whole brain volume, the activated network, and the pre-defined clusters. In the current study, we used the ICC toolbox to compute the test-retest reliability of the brain activation during the BART across two sessions in the whole brain volume, the activated network, and pre-defined ROIs. A priori ROIs in mesolimbic-frontal network were defined based on the BART activation results reported by Rao et al. (2008). Similar to a previous study (Van Den Bulk et al., 2013), bilateral occipital cortices associated with visual stimuli processing were also included as the control region. The MNI coordinates of selected ROIs are listed in Table 1 and displayed in Supplemental Fig. S1. According to the results of win outcome activation (Fig. 3A–B), the anatomical bilateral putamen was selected as ROIs for win outcomes using WFU_Pickatlas toolbox AAL template (http://fmri.wfubmc.edu/software/PickAtlas). In order to investigate the potential effects of trial numbers and head motion artifacts on test-retest reliability, additional analyses were conducted and the results were reported in supplementary material (see Table S4 and Table S5).

2.7. Analysis for spatial consistency: overlap volumes

For the brain activation associated with increased risk levels, the spatial overlaps based on Dice similarity coefficient (DSC) (Dice, 1945; Sørensen, 1948) of the two sessions were computed in 3D single subject space for all ROIs, as well as whole brain using the binarized image volumes. For the DSC of ROIs, the applied threshold level ranged from FWE corrected $p < 0.05$ to uncorrected $p < 0.005$ at the whole brain level. In order to ensure comparability between sessions, threshold levels were kept constant within the respective subjects across two sessions. In addition, the DSCs of whole brain activation across two sessions were computed at two different thresholds: FWE corrected $p < 0.05$ and uncorrected $p < 0.001$. The sample size and applied threshold for each ROI are shown in Supplementary Table S3. Due to the small number of win and loss outcome trials and individual variability, the DSCs of activation associated with win and loss outcomes were not computed. The DSC was calculated according to Rombouts et al. (1997) and ranges between 0 and 1 (Nettekoven et al., 2018; Wilson et al., 2017). We used the following guidelines to evaluate the DSC of the current study: Low (0.00–0.19), Low-moderate (0.20–0.39), Moderate (0.40–0.59), Moderate-high (0.60–0.79), and High (0.80–1.00).

3. Results

3.1. Test-retest reliability of behavioral performances during the BART

Several indices were computed to represent the behavioral performance, including the average adjusted pump, the number of win trials, the number of loss trials, the win ratio of all trials, and the number of balloon inflations (i.e. number of trials for risk processing). As illustrated in Table 2, no significant difference between two sessions was found in participants' behavioral performances (all $p > 0.5$). Excellent ICCs were observed for all performance indices across the two sessions (ICC > 0.79).

3.2. Test-retest reliability of brain activation during the BART

3.2.1. Whole-brain activation during the BART—Whole-brain activation patterns across the two sessions are presented for all three events: brain activation associated with the increased risk level (see Fig. 2 & Table S1), brain activation associated with win outcomes (see Fig. 3A–B & Table S2), and brain activation associated with loss outcomes (see Fig. 3C–D & Table S2).

As shown in Fig. 2, we observed robust activation within the mesolimbic-frontal regions associated with increased risk levels, including the thalamus, bilateral striatum, bilateral anterior insula, bilateral dorsal lateral prefrontal cortex (DLPFC), and anterior cingulate/ medial frontal cortex (ACC/MFC) during both BART fMRI scan sessions. The occipital cortex was activated in response to the risk levels as well. As shown in Fig. 3, during the outcome processing phase, for both test sessions, we observed activation in bilateral putamen associated with win outcomes, bilateral anterior insula associated with loss outcomes, and visual areas for both outcomes. No significant differences in brain activation were found between the two scan sessions.

3.2.2. Test-retest reliability of brain activation—We computed the test-retest reliability of brain activation for three contrasts in the whole brain volume, the activated network, and the pre-defined ROIs. Results of the brain activation associated with the increased risk level are shown in Fig. 4. The ICC frequency distribution is shown in Fig. 4A. As expected, the activation test-retest reliability for the activated network (ICC = 0.62) was higher than that for the whole brain volume (ICC = 0.53). Although ICCs and t-scores are statistically independent, Fig. 4B shows higher ICCs within regions showing stronger activation. Additionally, we calculated the median ICC values in the whole brain volume, the activated network, and the pre-defined ROIs, which are shown in Fig. 4C.

Results of the brain activation in response to win outcomes and loss outcomes are shown in Fig. 5 and Fig. 6, respectively. For both win outcomes and loss outcomes, the activation test-retest reliability of the activated network was higher than that of the whole brain volume (Figs. 5A and 6A). The positive correlations between median ICC and t-scores were also observed for win and loss outcomes (Figs. 5B and 6B), indicating higher ICCs within regions showing stronger activation. In addition, the median ICC values for the whole brain volume, the activated network, and ROIs are shown in Figs. 5C and 6C. According to the brain activation patterns for outcomes (Fig. 3), we selected the bilateral putamen as ROIs for win outcomes and selected the bilateral insula as ROIs for loss outcomes. The detailed ROI-based ICCs for three contrasts are reported in Table 3.

According to the ICCs standard applied in the previous fMRI studies (Brandt et al., 2013; Fournier et al., 2014), we used the following guidelines to evaluate the test-retest reliability of the current study: Poor (ICC < 0.40), Fair (0.40 ≤ ICC < 0.60), Good (0.60 ≤ ICC < 0.75), and Excellent (ICC ≥ 0.75). Based on this standard, our results demonstrated that the brain activation associated with the increased risk level during the BART had a good test-retest reliability within the activated network (ICC = 0.62) and a fair reliability in the whole brain volume (ICC = 0.53). As shown in Table 3, for a priori ROIs, the ICCs for the activation in response to risk levels ranged from fair (R. insula, $\text{ICC}_{\text{med}} = 0.51$) to good (RDLPFC, $\text{ICC}_{\text{med}} = 0.71$). Specifically, except for the right insula, the activation in all selected ROIs had good reliabilities (ICC > 0.60). In contrast, the test-retest reliabilities of the brain activation in response to win and loss outcomes were not as good as those of the activation associated with increased risk levels. The median ICCs were poor in the whole brain volumes for both outcomes (ICCs < 0.40), while in the activated network, the ICCs were fair for loss outcomes (ICCs = 0.42) but still poor for win outcomes (ICCs = 0.33). For the ROI-based ICCs, both outcomes had fair test-retest reliabilities (ICCs > 0.40) (Table 3).

3.2.3. Overlaps between ICC maps and group t-maps—In order to examine the relationships between the brain activation strength and the test-retest reliability, we first generated three ICC maps for each contrast using the ICC toolbox (Caceres et al., 2009) and extracted three group t-maps from the first scan session image data for each contrast. Then we overlapped the ICC maps and the group t-maps together, as shown in Fig. 7 (Red: group t-map; Green: ICC map; Yellow: overlapping regions). Threshold of t-values corresponding to p = 0.001 was applied to generate the activation group t-maps. ICC values were thresholded at 0.5. Overall, the ICC maps and the group t-maps showed fairly good overlap in the brain activation associated with risk levels (Fig. 7. Top), including the ACC/MFC, thalamus, bilateral striatum, right DLPFC, and occipital cortex, which indicated high ICC and high activation t-values. In contrast, the overlaps between ICC maps and the group t-maps for both win and loss outcomes were relatively poor. However, the high ICC regions in the ICC maps were close to the high t-value regions in the group t-maps, as shown in the bilateral putamen for the win outcomes and the bilateral insula for the loss outcomes.

3.3. Spatial consistency: overlap volumes

In order to investigate the spatial consistency of the activation in response to increased risk level in the main regions and the whole brain volume across the two scan sessions, calculation of Dice’s similarity coefficient (DSC) was conducted at the individual level. As shown in Fig. 8, most ROIs had a moderate spatial consistency level (0.40 < DSC < 0.59), and the bilateral insula had a moderate-high spatial consistency level (0.60 < DSC < 0.65). Additionally, the DSCs of whole brain activation across the two scan sessions were computed at two different thresholds: uncorrected p < 0.001 and FWE corrected p < 0.05. Results showed lower overlap volumes for a more stringent threshold level (0.34 ± 0.28, FWE corrected p < 0.05) than for a more liberal threshold level (0.50 ± 0.25, uncorrected p < 0.001).

4. Discussion

Despite the importance of the reproducibility of fMRI results, the reliability of brain activation during the widely used BART paradigm has not been thoroughly evaluated. To our knowledge, the current study is the first to investigate the test-retest reliability of brain activity during the BART. Specifically, we used the same scanner, applied the same image acquisition sequence and parameters, and conducted two scan sessions at the same time-of-day for each individual subject to minimize potential factors that may influencing the reliability. Using the ICC toolbox (Caceres et al., 2009), the intraclass correlation coefficient analyses were performed to assess the reliability of brain activation within the whole brain volume, the activated network, and the pre-defined ROIs across the two fMRI BART scan sessions. Analyses revealed the following main findings. First, similar to previous studies (White et al., 2008; Xu et al., 2013), we replicated the high ICC results for behavioral performances on the BART. Second, at the group level, we observed robust and consistent brain activation patterns for both scan sessions, including activation within the mesolimbic-frontal network, which was associated with increased risk level, bilateral putamen activation, which was associated with win outcomes, and bilateral insula activation, which was associated with loss outcomes. These findings are in line with the previous BART fMRI studies (Pan et al., 2018; Rao et al., 2008, 2011). Third, using the ICC analysis and the Dice's similarity analysis, we found fair to good test-retest reliability and moderate to moderate-high spatial consistency of the brain activation in the whole brain and all pre-defined ROIs during risk processing. The test-retest reliability of brain activation in response to outcome processing was fair in the main ROIs (i.e. bilateral putamen for win outcomes; bilateral insula for loss outcomes), but was poor in the whole brain volume.

4.1. Test-retest reliability of whole brain and activated network

Using the within-subject ICC analysis on the individual level brain activation, we observed a good reliability for the activated network during the risk processing phase (median ICC = 0.62). Such reliability is lower than that of the BART behavioral performances (ICC range from 0.798 to 0.873, see Table 2). This is consistent with several previous test-retest fMRI studies using other tasks (Plichta et al., 2012; Upadhyay et al., 2015; Yang et al., 2019). Two possible reasons may explain why the ICC values of fMRI results are in general not comparable to those of behavioral measures (Bennett and Miller, 2010; Vul et al., 2009). First, task-induced brain activation and behavioral performance are influenced by different factors. The main factors contributing to the BART performance are cognitive elements, including arousal, cognitive strategies, learning, and personal risk propensity. In contrast, the reproducibility of fMRI is not only subject to the cognitive factors, but also to various non-psychological factors. These additional factors include variations in subjects' head position during scanning, the MR field inhomogeneity, the signal to noise ratio (SNR) of BOLD contrast, as well as the cardiac, respiratory, and motion artifacts. All of these can affect image reproducibility (McGonigle, 2012; Raemaekers et al., 2007; Veltman et al., 2000). Second, the ICCs were calculated based on the group-level activation threshold, while the individual-level activation threshold could vary dramatically (Nettekoven et al., 2018; Stevens et al., 2013). Thus, using a group-level threshold to predict single-subject brain activity might have introduced additional variance. Taken together, it is difficult to make direct comparisons between the ICC of fMRI results and the ICC of behavioral performances. In fact, it was demonstrated that most task-induced BOLD activation showed ICC values in a range of 0.33–0.66 (Bennett and Miller, 2010; Vul et al., 2009).

The test-retest reliability of brain activation in response to risk levels was higher than that of brain activation during the outcome and feedback process, in which the activated network reliability was relatively low, regardless of win or loss outcomes (median ICCs < 0.45). This difference might be due to the specific task design. The current BART version was modified to be compatible with an 8 min fMRI scan session and the average number of balloon trials subjects completed during each scan were about 21 balloons (session 1: 21.09 ± 4.84; session 2: 20.56 ± 3.93, see Table 2). The trial numbers for win and loss outcomes were even less, especially for loss trials which were only about 5 balloons (session 1: 5.09 ± 2.76; session 2: 5.00 ± 2.90, see Table 2). Previous studies have suggested that about 25 trials are needed to yield stable activation maps and provide sufficient SNR for the imaging results (Huettel and McCarthy, 2001; Murphy and Garavan, 2005). Therefore, the small number of outcome feedback trials might be one reason for its lower reliability observed in this study.

4.2. Test-retest reliability of ROI-based activation

Overall, the ROI-based ICCs of the activation associated with the increased risk level were fair to good, which is consistent with our expectation. Specifically, the ACC/MFC, right DLPFC, the thalamus, and the occipital cortex had better reliability than other regions, indicating that these regions can serve as stable biomarkers across fMRI scan sessions on the individual subject level. It is worth mentioning that we found fair to good reliability of the brain activation associated with the risk processing in the subcortical regions, including the bilateral striatum and bilateral insula. Measured fMRI activity generally has higher resolution on the surface of cerebral cortex than in subcortical regions. Subcortical fMRI must overcome two challenges: spatial resolution and physiological noise (Katyal et al., 2012; Maugeri et al., 2018). Therefore, the SNR in subcortical regions is usually worse than that in cortical areas, which might be one of the reasons that the test-retest reliability of the activation in the bilateral striatum and the bilateral insula was lower than that in the cortical regions. Thus, the fair to good reliability of the activation for the subcortical regions can be regarded as acceptable and may indicate that the mesolimbic pathway had relatively stable activation during the BART over time.

For the outcome phases, the ICC values of the activation in key regions were also fair, including the bilateral putamen activation associated with win outcomes and the bilateral insula activation associated with loss outcomes. As the control region, the occipital cortex showed good reliability in the risk condition, but only fair reliability in the outcome conditions. This could be related to the differences in number of trials and/or visual stimuli during risk and outcome conditions. Given the limited number of trials for the loss/win outcomes as compared to the large number of inflation trials for risk processing, the reduced reliability for the occipital activation in the outcome conditions are not surprising.

Many factors may affect MR image quality and impact the reliability results, including MRI acquisition sequence parameters, experimental designs, sample size, data analysis approaches, and individual differences in task performance. Thus, achieving high reliability from fMRI data is a challenging goal. Indeed, Bennett and Miller (2010) reviewed 62 fMRI test-retest studies and reported that most studies exhibited reliability ranging from 0.33 to 0.66 measured by the ICC values. The median ICC values found in this study are consistent with this range, yet our ICC values are lower when compared to the relatively high ICC values (>0.75) in previous studies using the probabilistic learning task (Aron et al., 2006; Freyer et al., 2009) and the reward anticipation task (Plichta et al., 2012). A potential explanation is the relatively short test-retest interval (one to three days) in this study as compared to the intervals in previous studies (two weeks to one year). Practice effects or habituation effects may remain with short test-retest intervals, which might influence the reliability. Previous studies have shown that practice or habituation effects may increase (Iacoboni et al., 1996; Kami et al., 1995) or decrease (Breiter et al., 1996; Phan et al., 2003) the brain activation during the second scan, therefore reduce the test-retest reliability. For instance, Chase et al. (2015) used a reward learning task and found lower ICC in the ventral striatum, suggesting a learning effect in brain activation with a short interval, while long intervals may reset potential habituation effects (Plichta et al., 2012). Moreover, the current study's the sample size (n = 34) is larger than the previous studies (n in the range of 8–25), which may affect the statistical power and influence the reliability. Studies have shown that reliability varies substantially with different statistical thresholds (Stevens et al., 2013). In addition, the test-retest reliability of brain responses to the BART is lower than the reliability of BART performance (ICC ranged from 0.798 to 0.896). These findings are in line with several studies suggesting that fMRI activation may be less stable than task performance (Plichta et al., 2012; Upadhyay et al., 2015; Yang et al., 2019) and highlight the importance of additional research in this area.

4.3. Spatial consistency

Using the Dice's similarity analysis, we explored the spatial consistency of the brain activation associated with increased risk levels. Our ROI-based data revealed moderate to moderately high Dice's similarity coefficients across the two fMRI scan sessions. Surprisingly, however, we observed relatively lower Dice's similarity in the ACC/MFC (0.46 ± 0.13) but higher Dice's similarity in bilateral insula (0.61 ± 0.10), which is in contrast with the ICC results pattern. One possible explanation for this finding is that the applied threshold and cluster size cut-off used might have influenced the results (Kauppi et al., 2017; Nettekoven et al., 2018). In addition, during the DSC analyses, we defined "activated" regions as ROIs at the single subject level by applying consistent threshold for each subject across both scans. However, there were remarkable inter-individual differences in activation intensity and we could not yield activation in some ROIs or for some subjects. Therefore, the number of subjects included in the DSC analysis was different for each ROI. All of these factors could affect the spatial consistency calculation. Additionally, we observed a moderate spatial consistency in the whole brain volume when using the same threshold of the group-level analyses (i.e. uncorrected $p < 0.001$), but the consistency was smaller when a more stringent threshold level was applied (i.e. FWE corrected $p < 0.05$). It should be noted that the whole brain-based analysis can lead to segmentation of both noise and stimulus-related regions (Kauppi et al., 2017); thus, performing and interpreting the whole-brain volume spatial consistency must be done with caution. Overall, in the current study, the spatial consistency pattern was in line with the test-retest reliability in most regions and suggested an acceptable level of the reproducibility of BART activation.

4.4. Limitations

There are a number of factors to consider with respect to the interpretation of our results. First, although the current BART paradigm is comparable with other BART neuroimaging studies (e.g., Claus and Hutchison, 2012; Galván et al., 2013; Helfinstein et al., 2014; Lei et al., 2017; Qu et al., 2015), it should be noted that our BART paradigm is a modified version of the original BART. For example, the maximum numbers of pumps in the current study is 12, but is 128 in the original BART. The wager size and the probability of explosion both monotonically increase in the current BART, whereas the probability of explosion follows a quasi-normal distribution in the original BART. Further, there is a cost to subjects (from their earnings) for exploded balloons in the current BART, but there is no such cost/loss of money from total earnings in the original BART. These different features may affect participants' risk-taking behavior and influence the generalizability of current findings. Future neuroimaging studies are needed to replicate our findings using the paradigms closer to the original BART.

Another consideration in the interpretation of our results is the duration of the BART scan was relatively short (8 min) and the numbers of both win and loss outcome trials were limited (both < 25). Consequently, our modified BART paradigm might be suboptimal for assessing the reliability of brain activation during outcome processing. Therefore, if a future study aims to examine brain activation in response to outcome processing, multiple BART sessions should be employed to obtain sufficient number of trials and more robust results. Future research may also compare fMRI data with different number of trials to determine the minimum trial number required to obtain stable and acceptable reliability for different conditions.

It is important to note that the two BART fMRI sessions in this study were conducted within a short interval of one to three days. As we mentioned before, the short interval might lead to practice effects from the first test session, which would influence the reliability. Future research should investigate the BART fMRI reliability over a longer interval and determine the relationship between different time intervals and the BART test-retest reliability. Finally, we did not collect the participants' IQ, socioeconomic status (SES), personal risk propensity, or other risk-taking related assessment in this study, which might also be related to the test-retest reliability. In future studies, these related factors should be taken into account.

5. Conclusions

In summary, we evaluated the test-retest reliability of brain activation patterns during repeated fMRI scans of the widely used BART paradigm in a cohort of 34 healthy adults in the present study. Our analyses showed that test-retest reliability of brain activation in response to the BART risk-taking is good and acceptable in the mesolimbic-frontal network, while the reliability of brain activation in response to the loss or win outcomes is fair. These results have implications for future utility of the BART in neuroimaging research. For instance, our findings suggest that the BART in fMRI can be used to identify individual differences in trait-like risk-taking behavior and brain responses among healthy subjects as well as examine abnormal risk-taking behavior and altered brain activation in clinical populations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by the grants from the National Institutes of Health (NIH) (R21 DA032022, R01 MH107571, R21 AG051981, R01 HL102119), National Natural Science Foundation of China (71942003, 71671115), Shanghai International Studies University Major Research Project (20171140020), and the Program for Professors of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (TP2016020 and GZ2019003). The funders had no role in the study design, data collection and analysis, data interpretation, writing of the manuscript, or the decision to submit the article for publication. We gratefully acknowledge Ellen Paul and Mengsen Zhang for their help in imaging data collection.

References

Aklin WM, Lejuez CW, Zvolensky MJ, Kahler CW, Gwadz M, 2005 Evaluation of behavioral measures of risk taking propensity with inner city adolescents. Behav. Res. Ther 43 (2), 215–228. 10.1016/j.brat.2003.12.007. [PubMed: 15629751]

Aron AR, Gluck MA, Poldrack RA, 2006 Long-term test-retest reliability of functional MRI in a classification learning task. Neuroimage 29 (3), 1000–1006. 10.1016/j.neuroimage.2005.08.010. [PubMed: 16139527]

othersAtri A, O'Brien JL, Sreenivasan A, Rastegar S, Salisbury S, DeLuca AN, Sperling RA, 2011 Test-retest reliability of memory task functional magnetic resonance imaging in Alzheimer disease clinical trials. Archives of neurology 68 (5), 599–606. [PubMed: 21555634]

Baldwin SA, Murray DM, Shadish WR, 2005 Empirically supported treatments or type I errors? Problems with the analysis of data from group-administered treatments. J. Consult. Clin. Psychol 73 (5), 924–935. 10.1037/0022-006X.73.5.924. [PubMed: 16287392]

Banducci AN, Felton JW, Dahne J, Ninnemann A, Lejuez CW, 2015 Maternal risk taking on the balloon analogue risk task as a prospective predictor of youth alcohol use escalation. Addict. Behav 49, 40–45. 10.1016/j.addbeh.2015.05.011. [PubMed: 26046400]

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

Blautzik J, Keeser D, Berman A, Paolini M, Kirsch V, Mueller S, Meindl T, 2013 Long-term test-retest reliability of resting-state networks in healthy elderly subjects and patients with amnestic mild cognitive impairment. Journal of Alzheimer's Disease 34 (3), 741–754.

othersBornovalova MA, Cashman-Rolls A, O'donnell JM, Ettinger K, Richards JB, Lejuez CW, 2009 Risk taking differences on a behavioral task as a function of potential reward/loss magnitude and individual differences in impulsivity and sensation seeking. Pharmacol. Biochem. Behav 93 (3), 258–262. [PubMed: 19041886]

Braams BR, van Duijvenvoorde ACK, Peper JS, Crone EA, 2015 Longitudinal changes in adolescent risk-taking: a comprehensive study of neural responses to rewards, pubertal development, and risk-taking behavior. J. Neurosci 35 (18), 7226–7238. 10.1523/JNEUROSCI.4764-14.2015. [PubMed: 25948271]

Brandt DJ, Sommer J, Krach S, Bedenbender J, Kircher T, Paulus FM, Jansen A, 2013 Test-retest reliability of fMRI brain activity during memory encoding. Front. Psychiatry 4 (12), 1–9. 10.3389/fpsyt.2013.00163. [PubMed: 23346060]

Braun U, Plichta MM, Esslinger C, Sauer C, Haddad L, Grimm O, Walter H, 2012 Test–retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. Neuroimage 59 (2), 1404–1412. [PubMed: 21888983]

Breiter HC, Etcoff NL, Whalen PJ, Kennedy WA, Rauch SL, Buckner RL, Rosen BR, 1996 Response and habituation of the human amygdala during visual processing of facial expression. Neuron 17 (5), 875–887. [PubMed: 8938120]

Caceres A, Hall DL, Zelaya FO, Williams SCR, Mehta MA, 2009 Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage 45 (3), 758–768. 10.1016/j.neuroimage.2008.12.035. [PubMed: 19166942]

Callicott JH, Egan MF, Mattay VS, Bertolino A, Bone AD, Verchinksi B, Weinberger DR, 2003 Abnormal fMRI response of the dorsolateral prefrontal cortex in cognitively intact siblings of patients with schizophrenia. Am. J. Psychiatry 160 (4), 709–719. [PubMed: 12668360]

Cao H, Plichta MM, Sch€afer A, Haddad L, Grimm O, Schneider M, Tost H, 2014 Test–retest reliability of fMRI-based graph theoretical properties during working memory, emotion processing, and resting state. Neuroimage 84, 888–900. [PubMed: 24055506]

othersChase HW, Fournier JC, Greenberg T, Almeida JR, Stiffler R, Zevallos CR, Adams P, 2015 Accounting for dynamic fluctuations across time when examining fMRI test-retest reliability: analysis of a reward paradigm in the EMBARC study. PloS one 10 (5), e0126326. [PubMed: 25961712]

Chen B, Xu T, Zhou C, Wang L, Yang N, Wang Z, Weng XC, 2015 Individual variability and test-retest reliability revealed by ten repeated resting-state brain scans over one month. PLoS One 10 (12), e0144963. [PubMed: 26714192]

Claus ED, Hutchison KE, 2012 Neural mechanisms of risk taking and relationships with hazardous drinking. Alcohol Clin. Exp. Res 36 (6), 408–416. 10.1111/j.1530-0277.2011.01694.x.

Dice LR, 1945 Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Dosenbach NUF, Nardos B, Cohen AL, Fair DA, Power JD, Church JA, Barnes KA, 2010 Prediction of individual brain maturity using fMRI. Science 329 (5997), 1358–1361. [PubMed: 20829489]

Dubois J, Adolphs R, 2016 Building a science of individual differences from fMRI. Trends Cogn. Sci 20 (6), 425–443. 10.1016/j.tics.2016.03.014. [PubMed: 27138646]

Fang Z, Jung WH, Korczykowski M, Luo L, Prehn K, Xu S, Rao H, 2017 Postconventional moral reasoning is associated with increased ventral striatal activity at rest and during task. Scientific reports 7 (1), 7105. [PubMed: 28769072]

Fournier JC, Chase HW, Almeida J, Phillips ML, 2014 Model specification and the reliability of fMRI results: implications for longitudinal neuroimaging studies in psychiatry. PLoS One 9 (8). 10.1371/journal.pone.0105169.

Freyer T, Valerius G, Kuelz AK, Speck O, Glauche V, Hull M, Voderholzer U, 2009 Test-retest reliability of event-related functional MRI in a probabilistic reversal learning task. Psychiatry Res. Neuroimaging 174 (1), 40–46. 10.1016/j.pscychresns.2009.03.003.

Galván A, Schonberg T, Mumford J, Kohno M, Poldrack RA, London ED, 2013 Greater risk sensitivity of dorsolateral prefrontal cortex in young smokers than in nonsmokers. Psychopharmacology 229 (2), 345–355. [PubMed: 23644912]

Geerligs L, Renken RJ, Saliasi E, Maurits NM, Lorist MM, 2014 A brain-wide study of age-related changes in functional connectivity. Cerebr. Cortex 25 (7), 1987–1999.

Gorgolewski KJ, Mendes N, Wilfling D, Wladimirow E, Gauthier CJ, Bonnen T, Smallwood J, 2015 A high resolution 7-Tesla resting-state fMRI test-retest dataset with cognitive and physiological measures. Scientific data 2, 140054. [PubMed: 25977805]

Gorgolewski KJ, Storkey AJ, Bastin ME, Whittle I, Pernet C, 2013 Single subject fMRI test-retest reliability metrics and confounding factors. Neuroimage 69, 231–243. 10.1016/j.neuroimage.2012.10.085. [PubMed: 23153967]

Guo CC, Kurth F, Zhou J, Mayer EA, Eickhoff SB, Kramer JH, Seeley WW, 2012 One-year test–retest reliability of intrinsic connectivity network fMRI in older adults. Neuroimage 61 (4), 1471–1483. [PubMed: 22446491]

Helfinstein SM, Schonberg T, Congdon E, Karlsgodt KH, Mumford JA, Sabb FW, Poldrack RA, 2014 Predicting risky choices from brain activity patterns. Proceedings of the National Academy of Sciences 111 (7), 2470–2475.

Hopko DR, Lejuez CW, Daughters SB, Aklin WM, Osborne A, Simmons BL, Strong DR, 2006 Construct validity of the balloon analogue risk task (BART): relationship with MDMA use by inner-city drug users in residential treatment. J. Psychopathol. Behav. Assess 28 (2), 95–101. 10.1007/s10862-006-7487-5.

Huettel SA, McCarthy G, 2001 The effects of single-trial averaging upon the spatial extent of fMRI activation. Neuroreport 12 (11), 2411–2416. 10.1097/00001756-200108080-00025. [PubMed: 11496120]

Iacoboni M, Woods RP, Mazziotta JC, 1996 Brain-behavior relationships: evidence from practice effects in spatial stimulus-response compatibility. J. Neurophysiol 76(1), 321–331. [PubMed: 8836228]

Kami A, Meyer G, Jezzard P, Adams MM, Turner R, Ungerleider LG, 1995 Functional MRI evidence for adult motor cortex plasticity during motor skill learning. Nature 377 (6545), 155. [PubMed: 7675082]

Kathleen Holmes M, Bearden CE, Barguil M, Fonseca M, Serap Monkul E, Nery FG, Glahn DC, 2009 Conceptualizing impulsivity and risk taking in bipolar disorder: importance of history of alcohol abuse. Bipolar disorders 11 (1), 33–40. [PubMed: 19133964]

Katyal S, Greene CA, Ress D, 2012 High-resolution functional magnetic resonance imaging methods for human midbrain. J. Vis. Exp (63).

Kauppi JP, Pajula J, Niemi J, Hari R, Tohka J, 2017 Functional brain segmentation using inter-subject correlation in fMRI. Hum. Brain Mapp 38 (5), 2643–2665. [PubMed: 28295803]

Ko CH, Hsiao S, Liu GC, Yen JY, Yang MJ, Yen CF, 2010 The characteristics of decision making, potential to take risks, and personality of college students with Internet addiction. Psychiatry Res. 175 (1–2), 121–125. [PubMed: 19962767]

Koo TK, Li MY, 2016 A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med 15 (2), 155–163. [PubMed: 27330520]

Lauriola M, Panno A, Levin IP, Lejuez CW, 2014 Individual differences in risky decision making: a meta-analysis of sensation seeking and impulsivity with the balloon analogue risk task. J. Behav. Decis. Mak 27 (1), 20–36.

Lee TM, Guo LG, Shi HZ, Li YZ, Luo YJ, Sung CY, Lee TM, 2009 Neural correlates of traditional Chinese medicine induced advantageous risk-taking decision making. Brain and cognition 71 (3), 354–361. [PubMed: 19679384]

Lei Y, Wang L, Chen P, Li Y, Han W, Ge M, Yang Z, 2017 Neural correlates of increased risk-taking propensity in sleep-deprived people along with a changing risk level. Brain imaging and behavior 11 (6), 1910–1921. [PubMed: 27975159]

Lejuez CW, Aklin W, Daughters S, Zvolensky M, Kahler C, Gwadz M, 2007 Reliability and validity of the youth version of the balloon analogue risk task (BART–Y) in the assessment of risk-taking behavior among inner-city adolescents. J. Clin. Child Adolesc. Psychol 36 (1), 106–111. [PubMed: 17206886]

Lejuez CW, Aklin WM, Bornovalova MA, Moolchan ET, 2005 Differences in risk-taking propensity across inner-city adolescent ever-and never-smokers. Nicotine Tob. Res 7 (1), 71–79. [PubMed: 15804679]

Lejuez CW, Aklin WM, Jones HA, Richards JB, Strong DR, Kahler CW, Read JP, 2003 The balloon analogue risk task (BART) differentiates smokers and nonsmokers. Exp. Clin. Psychopharmacol 11 (1), 26. [PubMed: 12622341]

Lejuez CW, Read JP, Kahler CW, Richards JB, Ramsey SE, Stuart GL, Brown RA, 2002 Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). Journal of Experimental Psychology: Applied 8 (2), 75. [PubMed: 12075692]

Li Z, Kadivar A, Pluta J, Dunlop J, Wang Z, 2012 Test–retest stability analysis of resting brain activity revealed by blood oxygen level-dependent functional MRI. J. Magn. Reson. Imaging 36 (2), 344–354. [PubMed: 22535702]

Liao XH, Xia MR, Xu T, Dai ZJ, Cao XY, Niu HJ, He Y, 2013 Functional brain hubs and their test–retest reliability: a multiband resting-state functional MRI study. Neuroimage 83, 969–982. [PubMed: 23899725]

Lighthall NR, Mather M, Gorlick MA, 2009 Acute stress increases sex differences in risk seeking in the Balloon Analogue Risk Task. PLoS One 4 (7). 10.1371/journal.pone.0006002.

MacLean RR, Pincus AL, Smyth JM, Geier CF, Wilson SJ, 2018 Extending the balloon analogue risk task to assess naturalistic risk taking via a mobile platform. J. Psychopathol. Behav. Assess 40 (1), 107–116. [PubMed: 30505069]

MacPherson L, Magidson JF, Reynolds EK, Kahler CW, Lejuez CW, 2010 Changes in sensation seeking and risk-taking propensity predict increases in alcohol use among early adolescents. Alcohol Clin. Exp. Res 34 (8), 1400–1408. [PubMed: 20491737]

Mannfolk P, Nilsson M, Hansson H, Ståhlberg F, Fransson P, Weibull A, Olsrud J, 2011 Can resting-state functional MRI serve as a complement to task-based mapping of sensorimotor function? A test–retest reliability study in healthy volunteers. Journal of Magnetic Resonance Imaging 34 (3), 511–517. [PubMed: 21761469]

Maugeri L, Moraschi M, Summers P, Favilla S, Mascali D, Cedola A, Fratini M, 2018 Assessing denoising strategies to increase signal to noise ratio in spinal cord and in brain cortical and subcortical regions. Journal of Instrumentation 13 (02), C02028.

McGonigle DJ, 2012 Test-retest reliability in fMRI: or how I learned to stop worrying and love the variability. Neuroimage 62 (2), 1116–1120. 10.1016/j.neuroimage.2012.01.023. [PubMed: 22261373]

McGraw KO, Wong SP, 1996 Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1 (1), 30.

Mishra S, Novakowski D, 2016 Personal relative deprivation and risk: an examination of individual differences in personality, attitudes, and behavioral outcomes. Personal. Individ. Differ 90, 22–26.

Morrison MA, Churchill NW, Cusimano MD, 2016 Reliability of task-based fMRI for preoperative Planning : a test-retest study in brain tumor patients and healthy controls. Plos One 1–26. 10.6084/m9.figshare.2068371. Februari.

Murphy K, Garavan H, 2005 Deriving the optimal number of events for an event-related fMRI study based on the spatial extent of activation. Neuroimage 27 (4), 771–777. 10.1016/j.neuroimage.2005.05.007. [PubMed: 15961321]

Murray DM, Hannan PJ, Baker WL, 1996 A Monte Carlo study of alternative responses to intraclass correlation in community trials: is it ever possible to avoid Cornfield’s penalties? Eval. Rev 20 (3), 313–337. 10.1177/0193841X9602000305. [PubMed: 10182207]

Nettekoven C, Reck N, Goldbrunner R, Grefkes C, Weiß Lucas C, 2018 Short- and long-term reliability of language fMRI. Neuroimage 176, 215–225. 10.1016/j.neuroimage.2018.04.050. 10 2017. [PubMed: 29704615]

Pan Y, Lai F, Fang Z, Xu S, Gao L, Robertson DC, Rao H, 2018 Risk choice and emotional experience: a multi-level comparison between active and passive decision-making. J. Risk Res 1–28. [PubMed: 29348731]

Parkinson B, Phiri N, Simons G, 2012 Bursting with anxiety: adult social referencing in an interpersonal balloon analogue risk task (BART). Emotion 12 (4), 817. [PubMed: 22251046]

Phan KL, Liberzon I, Welsh RC, Britton JC, Taylor SF, 2003 Habituation of rostral anterior cingulate cortex to repeated emotionally salient pictures. Neuropsychopharmacology 28 (7), 1344–1350. 10.1038/sj.npp.1300186. [PubMed: 12784119]

Plichta MM, Schwarz AJ, Grimm O, Morgen K, Mier D, Haddad L, Colman P, 2012 Test–retest reliability of evoked BOLD signals from a cognitive–emotive fMRI test battery. Neuroimage 60 (3), 1746–1758. [PubMed: 22330316]

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

Qi X, Du X, Yang Y, Du G, Gao P, Zhang Y, Zhang Q, 2015 Decreased modulation by the risk level on the brain activation during decision making in adolescents with internet gaming disorder. Frontiers in behavioral neuroscience 9, 296. [PubMed: 26578922]

Qu Y, Galvan A, Fuligni AJ, Lieberman MD, Telzer EH, 2015 Longitudinal changes in prefrontal cortex activation underlie declines in adolescent risk taking. J. Neurosci 35 (32), 11308–11314. 10.1523/JNEUROSCI.1553-15.2015. [PubMed: 26269638]

Raemaekers M, Du Plessis S, Ramsey NF, Weusten JMH, Vink M, 2012 Test-retest variability underlying fMRI measurements. Neuroimage 60 (1), 717–727. 10.1016/j.neuroimage.2011.11.061. [PubMed: 22155027]

Raemaekers M, Vink M, Zandbelt B, van Wezel RJA, Kahn RS, Ramsey NF, 2007 Test-retest reliability of fMRI activation during prosaccades and antisaccades. Neuroimage 36 (3), 532–542. 10.1016/j.neuroimage.2007.03.061. [PubMed: 17499525]

Raemaekers Mathijs, Ramsey NF, Vink M, van den Heuvel MP, Kahn RS, 2006 Brain activation during antisaccades in unaffected relatives of schizophrenic patients. Biol. Psychiatry 59 (6), 530–535. [PubMed: 16165103]

Rao H, Korczykowski M, Pluta J, Hoang A, Detre JA, 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. [PubMed: 18582578]

Rao H, Mamikonyan E, Detre JA, Siderowf AD, Stern MB, Potenza MN, Weintraub D, 2011 Decreased ventral striatal activity with impulse control disorders in. Parkinson's Dis. 25 (11), 1660–1669. 10.1002/mds.23147.Decreased.

Reynolds EK, MacPherson L, Schwartz S, Fox NA, Lejuez CW, 2014 Analogue study of peer influence on risk-taking behavior in older adolescents. Prev. Sci 15 (6), 842–849. [PubMed: 24122411]

Rombouts SA, Barkhof F, Hoogenraad FG, Sprenger M, Valk J, Scheltens P, 1997 Test-retest analysis with functional MR of the activated area in the human visual cortex. Am. J. Neuroradiol 18 (7), 1317–1322. [PubMed: 9282862]

Sauder CL, Hajcak G, Angstadt M, Phan KL, 2013 Test-retest reliability of amygdala response to emotional faces. Psychophysiology 50 (11), 1147–1156. [PubMed: 24128307]

Schwartz DL, Tagge I, Powers K, Ahn S, Bakshi R, Calabresi PA, Papinutto N, 2019 Multisite reliability and repeatability of an advanced brain MRI protocol. Journal of Magnetic Resonance Imaging.

Shrout PE, Fleiss JL, 1979 Intraclass correlations: uses in assessing rater reliability.1. Shrout PE, Fleiss JL: intraclass correlations: uses in assessing rater reliability, 1979, 86:420–8 Psychological Bulletin Psychol. Bull 86 (2), 420–428. Retrieved from. http://www.ncbi.nlm.nih.gov/pubmed/18839484. [PubMed: 18839484]

Smith SM, Nichols TE, Vidaurre D, Winkler AM, Behrens TE, Glasser MF, Miller KL, 2015 A positive-negative mode of population covariation links brain connectivity, demographics and behavior. Nature neuroscienc 18 (11), 1565.

Somandepalli K, Kelly C, Reiss PT, Zuo XN, Craddock RC, Yan CG, Di Martino A, 2015 Short-term test–retest reliability of resting state fMRI metrics in children with and without attention-deficit/hyperactivity disorder. Developmental Cognitive Neuroscience 15, 83–93. [PubMed: 26365788]

Song J, Desphande AS, Meier TB, Tudorascu DL, Vergun S, Nair VA, Prabhakaran V, 2012 Age-related differences in test-retest reliability in resting-state brain functional connectivity. PLoS One 7 (12), e49847. [PubMed: 23227153]

Sørensen T, 1948 A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol. Skr 5, 1–34.

Specht K, Willmes K, Shah NJ, Jäncke L, 2003 Assessment of reliability in functional imaging studies. J. Magn. Reson. Imaging: Off. J. Int. Soc. Mag. Reson. Med 17 (4), 463–471.

Stevens MTR, D'Arcy RCN, Stroink G, Clarke DB, Beyea SD, 2013 Thresholds in fMRI studies: reliable for single subjects? J. Neurosci. Methods 219 (2), 312–323. 10.1016/j.jneumeth.2013.08.005. [PubMed: 23958749]

Upadhyay J, Lemme J, Anderson J, Bleakman D, Large T, Evelhoch JL, Becerra L, 2015 Test–retest reliability of evoked heat stimulation BOLD fMRI. J. Neurosci. Methods 253, 38–46. [PubMed: 26072245]

Van Den Bulk BG, Koolschijn PCM, Meens PH, Van Lang ND, Van Der Wee NJ, Rombouts SA, Crone EA, 2013 How stable is activation in the amygdala and prefrontal cortex in adolescence? A study of emotional face processing across three measurements. Developmental Cognitive Neuroscience 4, 65–76. [PubMed: 23043904]

van den Heuvel MP, Stam CJ, Kahn RS, Pol HEH, 2009 Efficiency of functional brain networks and intellectual performance. J. Neurosci 29 (23), 7619–7624. [PubMed: 19515930]

Veltman DJ, Friston KJ, Sanders G, Price CJ, 2000 Regionally specific sensitivity differences in fMRI and PET: where do they come from? Neuroimage 11 (6), 575–588. [PubMed: 10860787]

Vink M, Ramsey NF, Raemaekers M, Kahn RS, 2006 Striatal dysfunction in schizophrenia and unaffected relatives. Biol. Psychiatry 60 (1), 32–39. [PubMed: 16603134]

Vul E, Harris C, Winkielman P, Pashler H, Nichols T, Poline J-B, 2009 Reply to comments on puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect. Psychol. Sci 4 (3), 319–324. [PubMed: 26158970]

Weafer J, Baggott MJ, de Wit H, 2013 Test–retest reliability of behavioral measures of impulsive choice, impulsive action, and inattention. Exp. Clin. Psychopharmacol 21 (6), 475. [PubMed: 24099351]

White TL, Lejuez CW, de Wit H, 2008 Test-retest characteristics of the balloon analogue risk task (BART). Exp. Clin. Psychopharmacol 16 (6), 565–570. 10.1037/a0014083. [PubMed: 19086777]

Wilson SM, Bautista A, Yen M, Lauderdale S, Eriksson DK, 2017 Validity and reliability of four language mapping paradigms. Neuroimage: Clin. 16, 399–408. [PubMed: 28879081]

Xu S, Korczykowski M, Zhu S, Rao H, 2013 Assessment of risk-taking and impulsive behaviors: a comparison between three tasks. Soc. Behav. Personal 41 (3), 477–486. 10.2224/sbp.2013.41.3.477.

Xu S, Pan Y, Qu Z, Fang Z, Yang Z, Yang F, Rao H, 2018 Differential effects of real versus hypothetical monetary reward magnitude on risk-taking behavior and brain activity. Scientific reports 8 (1), 3712. [PubMed: 29487303]

Xu S, Pan Y, Wang Y, Spaeth AM, Qu Z, Rao H, 2016 Real and hypothetical monetary rewards modulate risk taking in the brain. Sci. Rep 6, srep29520.

Yang FN, Xu S, Spaeth A, Galli O, Zhao K, Fang Z, Rao H, 2019 Test-retest reliability of cerebral blood flow for assessing brain function at rest and during a vigilance task. NeuroImage 193, 157–166. [PubMed: 30894335]

Yarkoni T, 2015 Neurobiological substrates of personality: a critical overview. APA Handb. Person. Soc. Psychol 4, 61–83.

Zuo XN, Xing XX, 2014 Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: a systems neuroscience perspective. Neurosci. Biobehav. Rev 45, 100–118. 10.1016/j.neubiorev.2014.05.009. [PubMed: 24875392]

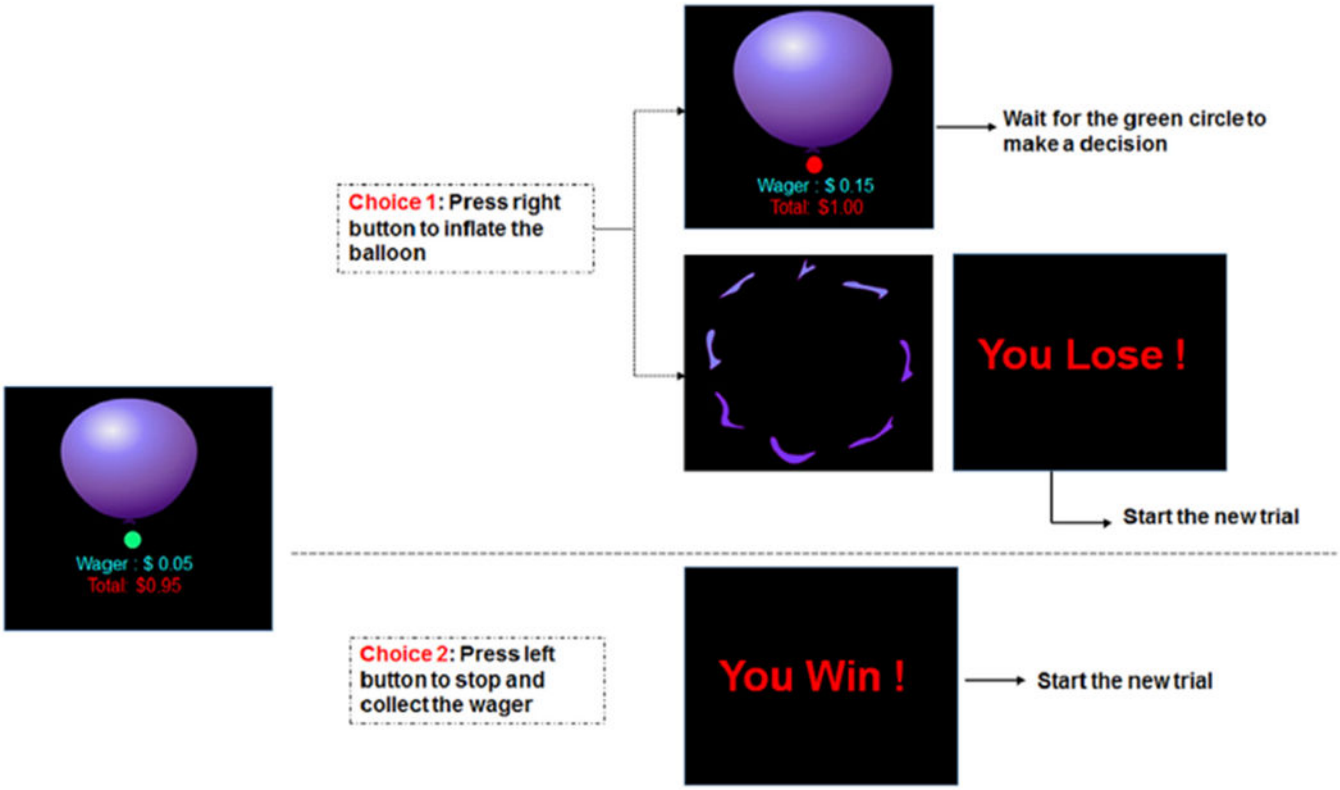

Choice 1: Press right button to inflate the balloon
Wager : $ 0.15
Total: $1.00
Wait for the green circle to make a decision
You Lose !
Start the new trial
Wager : $ 0.05
Total: $0.95
Choice 2: Press left button to stop and collect the wager
You Win !
Start the new trial

Fig. 1.

A sample BART trial showing inflation risks and outcome feedbacks. Participants could choose to continue or discontinue inflating the balloon at each turn.

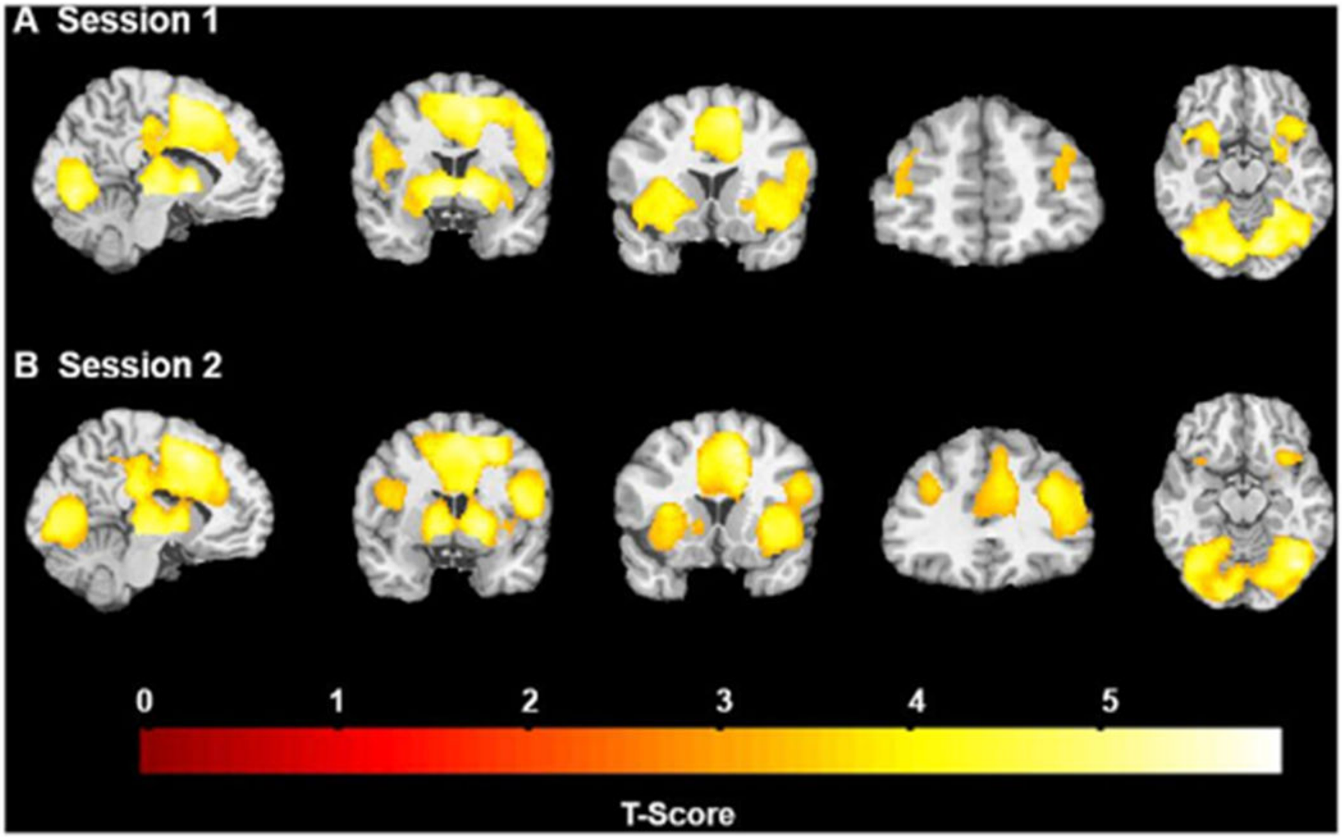


Fig. 2.

Whole-brain activation associated with the increased risk level in (A) session 1 and (B) session 2. Statistical inferences were performed at a threshold of uncorrected $p < 0.001$ at the whole-brain level and $p < 0.05$, family wise error (FWE) corrected at the cluster level.

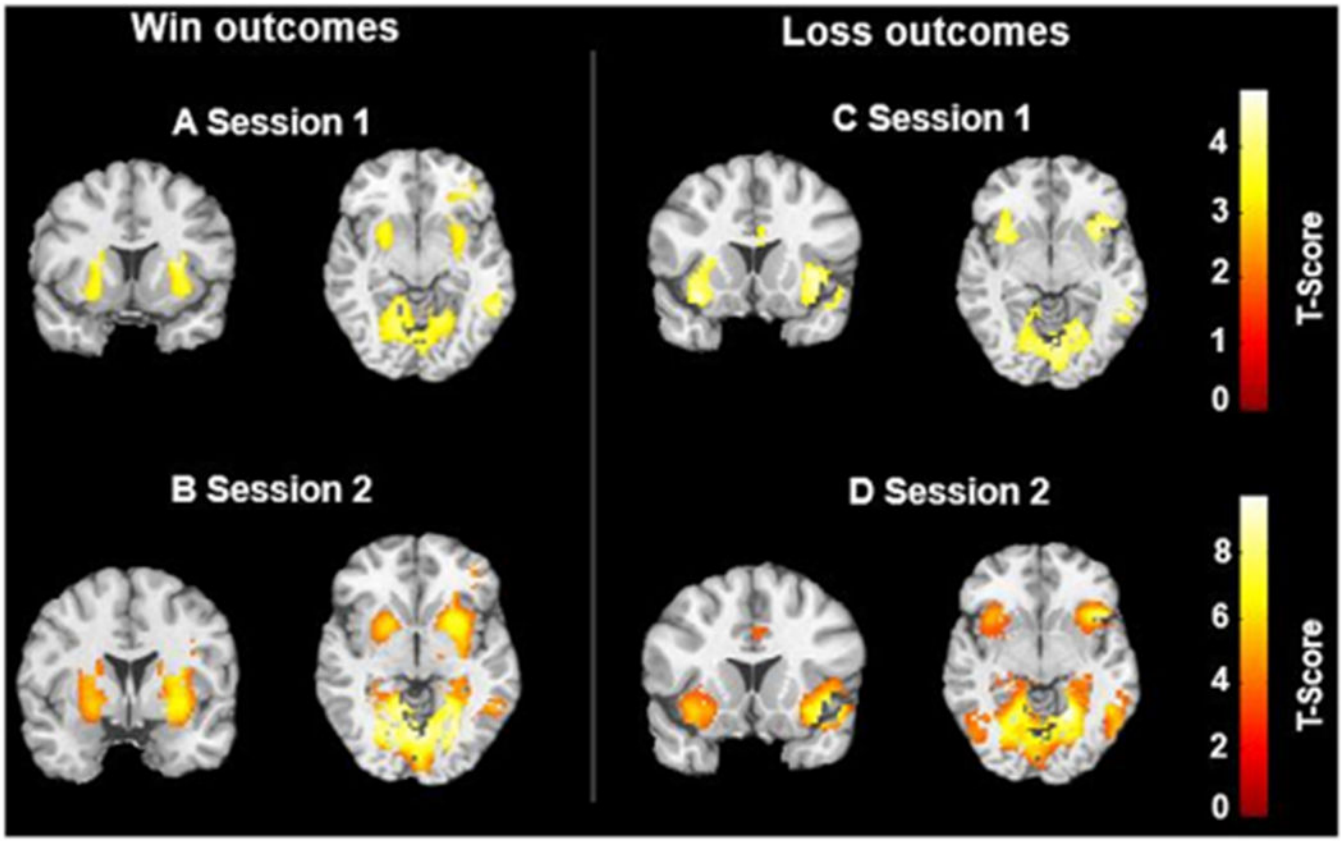


Fig. 3. Whole-brain activation in response to win (A–B) and loss (C–D) outcomes across two test sessions.

Statistical inferences were performed at a threshold of uncorrected $p < 0.001$ at the whole-brain level.

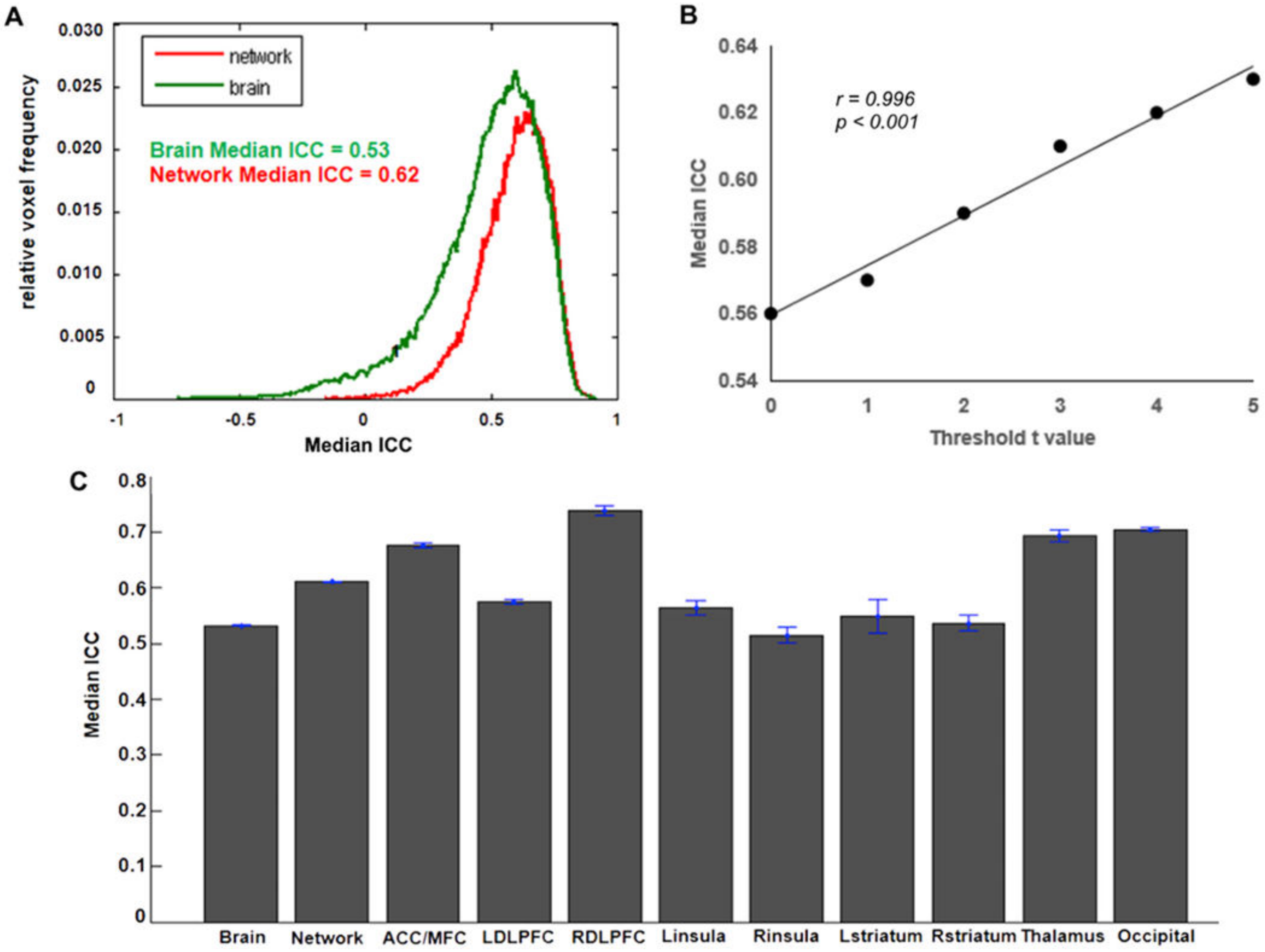
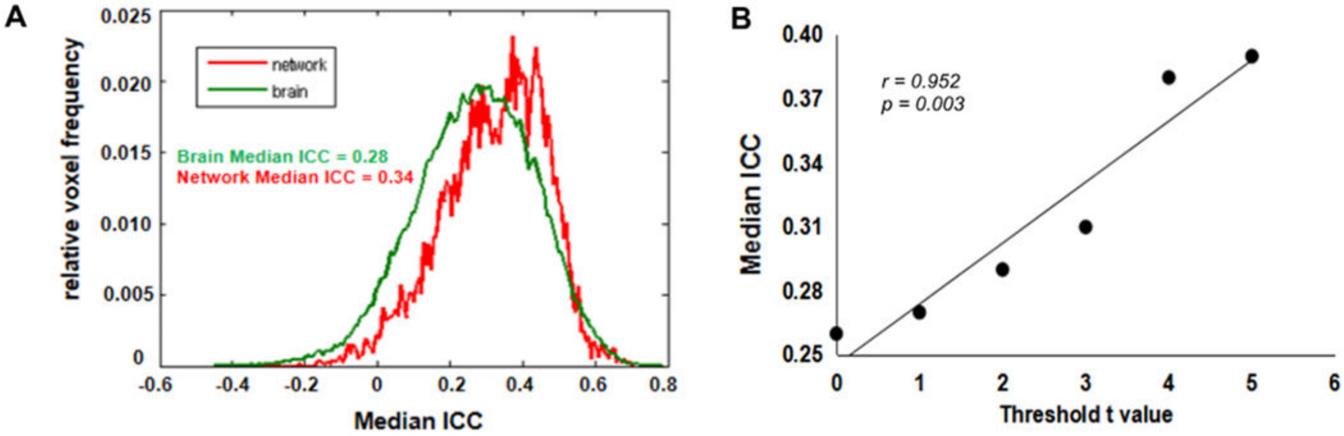


Fig. 4. Test-retest reliability of BART activation associated with increased risk level

(A) ICC frequency distribution for the whole-brain (Green line) and for the voxels in the activated network (Red line). The brain activation results from the first test session were used to define the “activated network” here. Voxels were classified as active if they had t-values t > 3.36 (corresponding to p < 0.001) (B) Correlation between median ICC and threshold t-scores. (C) Median ICC values for the whole-brain, activated network, and ROIs.



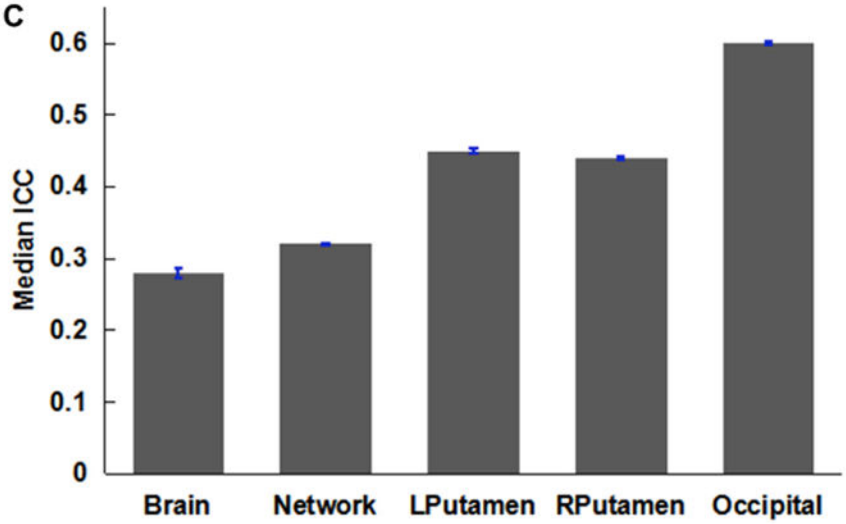


Fig. 5. Test-retest reliability of BART activation associated with win outcomes.

(A) ICC frequency distribution for the whole-brain (Green line) and for the voxels in the activated network (Red line). The brain activation results from the first test session were used to define the “activated network” here. Voxels were classified as active if they had t-values t > 3.36 (corresponding to p < 0.001). (B) Positive correlation between median ICC and threshold t-scores. (C) Median ICC values for the whole-brain, activated network, and ROIs.

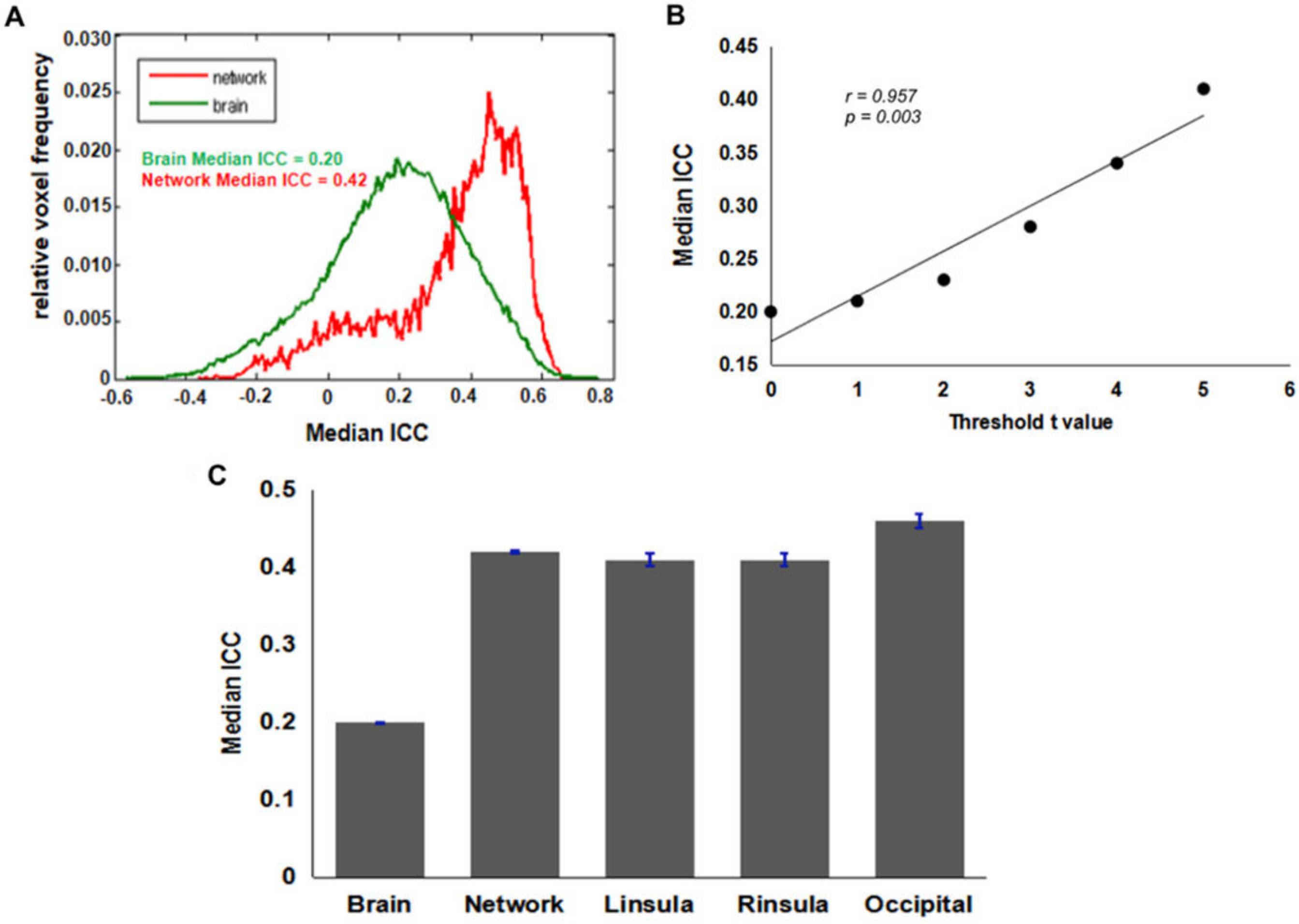


Fig. 6. Test-retest reliability of BART activation associated with loss outcomes

(A) ICC frequency distribution for the whole-brain (Green line) and for the voxels in the activated network (Red line). The brain activation results from the first test session were used to define the "activated network" here. Voxels were classified as active if they had t-values t > 3.36 (corresponding to p < 0.001). (B) Positive correlation between median ICC and threshold t-scores. (C) Median ICC values for the whole-brain, activated network, and ROIs.

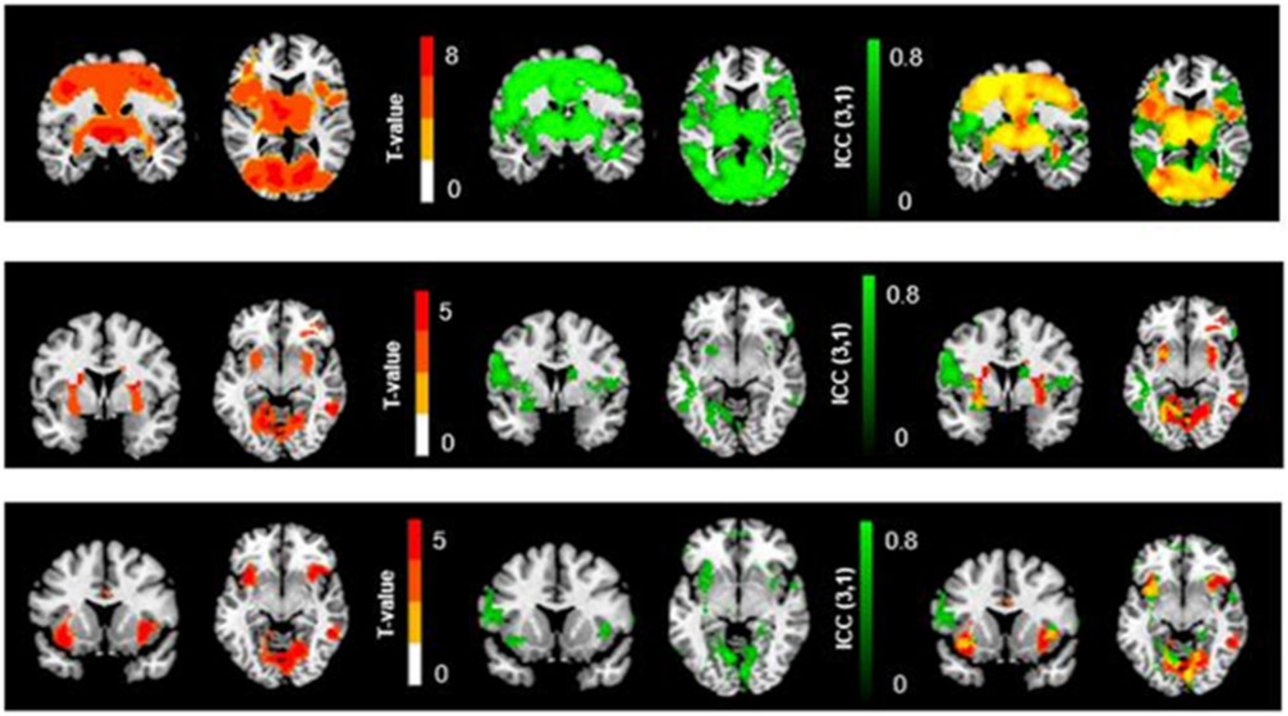


Fig. 7. Overlaps between ICC maps and group t-maps.

Top: Brain activation associated with increased risk levels. Middle: Brain activation in response to win outcomes. Bottom: Brain activation in response to loss outcomes. Left-Red: Group t-maps (t > 3.36 for risk; t > 3.36 for win; t > 3.36 for loss. Thresholding of t-values corresponding to p = 0.001). Middle-Green: ICC $_{(3,1)}$ maps (ICC > 0.5). Right-Yellow: Overlapping regions between group t-maps and ICC maps.

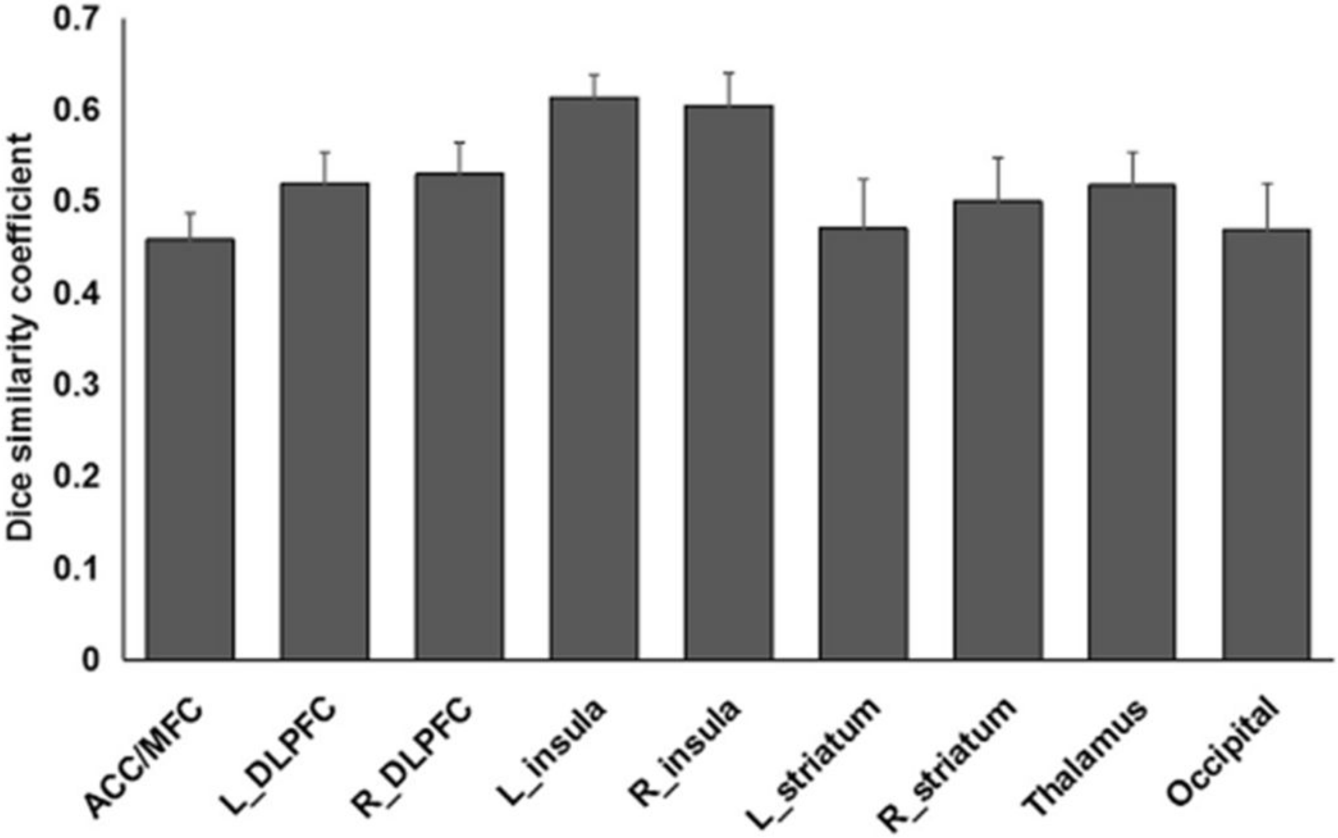


Fig. 8.

The average Dice's similarity coefficients of ROIs in mesolimbic-frontal and occipital cortex across two sessions at the individual level.

Table 1

Selected ROIs and their coordinates based on Rao et al. (2008).

Regions	Coordinates		
	x	y	z
ACC/MFC	0	12	42
Left DLPFC	−34	46	46
Right DLPFC	30	36	20
Left thalamus	−6	−12	−4
Right thalamus	6	−16	−2
Left striatum	−10	2	4
Right striatum	14	2	−2
Left insula	−34	18	−6
Right insula	38	10	−2
Left occipital	−28	−80	20
Right occipital	30	−76	24

Table 2

BART behavioral performances and correlations across session 1 and 2.

	Descriptive statistics			**Session 1 & 2 Correlations**			
	Session1 (M±SD)	**Session2 (M±SD)**	**p-value**	**Pearson r**	**ICC**	**ICC 95% CI**	
						Lower	**Upper**
Adjusted pump count	5.87 ± 1.39	5.90 ± 1.55	0.85	0.717 ***	0.836 ***	0.671	0.918
Win number	16.53 ± 5.58	16.44 ± 6.44	0.90	0.778 ***	0.873 ***	0.745	0.937
Loss number	5.09 ± 2.76	5.00 ± 2.90	0.83	0.659 ***	0.798 ***	0.594	0.900
Win ratio	0.75 ± 0.14	0.75 ± 0.16	0.87	0.698 ***	0.819 ***	0.637	0.910
Risk Trial number	138.56 ± 24.75	140.56 ± 27.14	0.33	0.900 ***	0.896 ***	0.802	0.947

Note: Adjust pump count refers to the average number of inflations for the win trials.

*** $p < 0.001$.

Table 3

ROI-based ICCs for three contrasts.

Condition	ROIs	ICC (3,1) Med	95% CI	ICC (3,1) Mean	95% CI
Risk	ACC/MFC	0.68	0.46–0.83	0.71	0.50–0.85
	L DLPFC	0.61	0.35–0.78	0.61	0.34–0.78
	R DLPDC	0.71	0.50–0.85	0.72	0.51–0.85
	R insula	0.51	0.21–0.72	0.50	0.20–0.72
	L striatum	0.62	0.36–0.79	0.57	0.35–0.79
	R striatum	0.60	0.33–0.78	0.64	0.39–0.80
	Thalamus	0.69	0.46–0.83	0.69	0.46–0.83
	Occipital	0.71	0.49–0.84	0.68	0.44–0.83
Win	L putamen	0.45	0.14–0.68	0.45	0.14–0.68
	R putamen	0.44	0.12–0.66	0.43	0.10–0.64
	Occipital	0.59	0.14–0.70	0.48	0.08–0.66
Loss	L insula	0.41	0.08–0.65	0.39	0.07–0.65
	R insula	0.41	0.09–0.65	0.43	0.11–0.67
	Occipital	0.46	0.05–0.64	0.41	0.02–0.62