Elucidating the role of the posterior medial frontal cortex in social conflict processing

Stephanie J. Wake[1], Ryuta Aoki[2], Kiyoshi Nakahara[3] & Keise Izuma[1,]*

[1] Department of Psychology, University of York, Heslington, York, YO10 5DD, UK.
[2] Research Institute for Future Design, Kochi University of Technology, Eikokuji, Kochi 780-8515, Japan
[3] Research Center for Brain Communication, Kochi University of Technology, Kami, Kochi 782-8502, Japan.

* Correspondence should be addressed to Keise Izuma
Current address
Department of Psychology, University of Southampton,
University Road, Southampton, SO17 1BJ, UK.
Email: K.Izuma@soton.ac.uk

ABSTRACT

A fundamental function of the brain is learning via new information. Studies investigating the neural basis of information-based learning processes indicate an important role played by the posterior medial frontal cortex (pMFC) in representing conflict between an individual's expectation and new information. However, specific function of the pMFC in this process remains relatively indistinct. Particularly, it’s unclear whether the pMFC plays a role in the detection of *conflict* of incoming information, or the *update* of their belief after new information is provided. In an fMRI scanner, twenty-eight Japanese students viewed scenarios depicting various pro-social/anti-social behaviors. Participants rated how likely Japanese and South Korean students would perform each behavior, followed by feedback of the actual likelihood. They were then asked to rerate the scenarios after the fMRI session. Participants updated their second estimates based on feedback, with estimate changes more pronounced for favorable feedback (e.g., higher likelihood of pro-social behavior than expected) despite nationality, indicating participants were willing to view other people favorably. The fMRI results demonstrated activity in a part of the pMFC, the dorsomedial prefrontal cortex (dmPFC), was correlated with social conflict (difference between participant's estimate and actual likelihood), but not the corresponding belief update. Importantly, activity in a different part within the dmPFC was more sensitive to *unfavorable* trials compared to favorable trials. These results indicate sensitivity in the pMFC (at least within the dmPFC) relates to conflict between desirable outcomes versus reality, as opposed to the associated update of belief.

Key Words: pMFC, social attitudes, favorability bias, conformity, learning

1. INTRODUCTION

Procuring knowledge via new information is one of the most important functions of the brain. We update our beliefs, knowledge and/or attitudes based on semantic factual information (e.g., how likely you are to become ill) as well as what other people think (i.e., social conformity). A number of past neuroimaging studies have investigated the neural mechanisms behind information-based learning processes, and currently available evidence converge to indicate an important role played by the posterior part of the medial frontal cortex (pMFC), particularly the dorsomedial prefrontal cortex (dmPFC) and dorsal anterior cingulate cortex (dACC), in representing the conflict between an individual's expectation and new information.

The pMFC is known to play a key role in processing reward prediction error (i.e., the difference between actual and predicted reward) in reinforcement learning tasks (specifically the ACC) (Sambrook and Goslin, 2015), and a number of neuroimaging studies have indicated that the pMFC plays a wider role, being involved in information-based learning in a variety of both social and non-social settings where there is no reward. For example, using a social conformity task, a functional magnetic resonance imaging (fMRI) study Klucharev et al., (2009) demonstrated that the rostral cingulate zone, a part of the ACC, tracked the discrepancy between individual's versus group's opinion so that the larger the conflict between one's and group's opinions, the higher the activity. This result has been replicated by other fMRI studies (Campbell-Meiklejohn et al., 2010; Izuma and Adolphs, 2013; Wu et al., 2016). Similarly, a number of electroencephalography (EEG) studies on social conformity (Chen et al., 2012; Huang et al., 2014; Kim et al., 2012; Schnuerch et al., 2014; Schnuerch and Gibbons, 2015; Shestakova et al., 2012) observed electrophysiological responses over the pMFC that track the conflict between one's versus group's opinion. The electrophysiological responses resemble the feedback-related negativity (FRN) signal, which is related to reward prediction error and is considered to be generated in the ACC (Holroyd and Coles, 2002; Sambrook and Goslin, 2015). Furthermore, more recently, Pine et al., (2018) demonstrated that the dmPFC, is involved in prediction error in learning based on semantic factual information.

Izuma and Adolphs (2013) further demonstrated that the pMFC doesn't simply represent the conflict between one's and others' opinion, but rather, it represents the conflict posed from desired versus undesired outcomes (Izuma, 2013). Izuma and Adolphs (2013) first replicated Klucharev et al.'s (2009) findings showing the pMFC (specifically the dmPFC) tracked the conflict between participant's and their fellow students' (participant's “liked” group) opinions. However, this pattern was completely reversed if it was an opinion of a “disliked” group; the pMFC activity was higher when their opinion was more *similar* to sex offenders' (disliked group) opinion. Thus, the results suggest the pMFC doesn't solely represent the distance between one's and others' opinion, but more embodies the divergence from desirable outcomes.

Although a number of studies have demonstrated that pMFC activity reflects the discrepancy between an individual's expectation (or opinion) and new information (or more broadly, the discrepancy between a desirable or ideal outcome, and reality), the exact roles of the pMFC in information-based learning still remains to be fully elucidated. More specifically, it remains unclear whether the pMFC plays a specific role in the detection of *conflict* of incoming information (with the dACC particularly involved in conflict monitoring and successive cognitive control; Mansouri et al., 2017; Shenhav et al., 2013), or is associated with the *update* of their belief after new information is provided. In previous studies, these two processes often co-occurred- making it difficult to disentangle them. For example, in a typical social conformity study, the larger the conflict between one's versus group's opinions, the more an individual conforms to the group's opinion (i.e., the greater update of their opinion).

Accordingly, the current study aimed to shed a new light on the role of the pMFC by utilizing cognitive bias, extending the findings of Izuma and Adolphs (2013). Numerous studies in psychology have demonstrated that we don't process information objectively, rather how we process new information is heavily affected by various cognitive biases. For example, as a general rule we tend to seek and formulate our attitudes based on information that already aligns with our own ideals, a phenomenon known as confirmation bias (Knobloch-Westerwick et al., 2015; Lord et al., 1979; Sunstein et al., 2016). Thus, how we update our belief depends on whether new information is consistent with how an individual already sees the world. Appropriately, by utilizing a cognitive bias, we can dissociate the level of conflict from the level of belief updating (e.g., the same degree of conflict can predict different levels of belief updating dependent on whether it is consistent with their preexisting ideals).

Confirmation bias here was elicited using an intergroup paradigm, specifically Japanese participants perceptions of other Japanese individuals (in-group) versus South Korean individuals (out-group), whom historically have a tense relationship (see Izuma et al., 2019; Lee, 1985). The vast social body of research regarding inter-group relations informs us that general favoritism towards the in-group and derogation towards an out-group tends to be a common nature of human group behaviour (for example Tajfel, 1982; Tajfel, 2010). Extensions to neuroscience research have been made increasingly apparent (for a recent review see Molenberghs and Louis, 2018; Hackel et al., 2017). A recent example comes from Lin et al., (2018), who found that after participants rated emotional stimuli in the scanner, they were more likely to change their evaluations to be more similar to the evaluations other in-group members made compared to the out-group. This shift was tracked by neural activity in the ventral striatum, dmPFC, mPFC, posterior superior temporal sulcus (pSTS), temporal pole, amygdala and insula (see also Huang et al. 2019). Thus, we applied an inter-group context to promote confirmation bias, directly manipulating the level of bias participants are presented with.

In the study, Japanese university students viewed a series of scenarios which describe either a pro-social or anti-social behavior inside an MRI scanner. Their task was to estimate how typical Japanese and South Korean students answered a series of questions relating to how they would respond in said scenarios (Figure 1). After they gave their rating, participants were presented with the rating given by Japanese or South Korean students (i.e., what percentage of Japanese or South Korean students were willing to perform the pro- or anti-social behavior). After participants had gone through all scenarios and feedback, they were then asked to rerate the scenarios as an experimental task outside of the scanner to index the level of belief updating.

Behaviorally, we expected that how much individuals updated their belief about Japanese and South Korean students depends on their attitudes toward Japan and South Korea, respectively, and the prosocial nature of the feedback presented. To the extent that our Japanese participants have positive attitudes toward Japan, they would update their belief about Japanese students more if new information allows them to see other Japanese students more favorably (e.g., if more Japanese students were willing to perform a pro-social behavior than expected). We expected a similar pattern for the South Korea condition, but this favorability bias would be less pronounced because of participants' less positive attitudes toward South Korea (outgroup) compared to Japan (ingroup) (i.e., participants’ would be more willing to view ingroup members favorably compared to outgroup members).

Furthermore, the study aimed to test the two competing hypotheses regarding pMFC activity, specifically the dmPFC. First, if the dmPFC encodes the conflict between a desirable state versus reality, its activity should be more sensitive to the difference between one's estimate and actual feedback when the feedback is in an *unfavorable* direction (conflict hypothesis). In contrast, if the pMFC plays a role in belief updating, its activity should be more sensitive to the difference when the feedback is in a *favorable* direction where we expected a larger update of their belief (update hypothesis).

2. METHOD

2.1. Participants

Twenty-nine right-handed Japanese students with no psychiatric history were recruited via a participant pool at the Kochi University of Technology. One participant was excluded from the analysis due to excessive head motion (i.e., >3mm). The final sample consists of 28 participants (Male = 16, Female = 12; mean age = 20.3). Note that due to a technical fault with the scanner, for one subject, fMRI data after 6 minutes of the first session were not obtained. Accordingly, for the first session of this subject, the fMRI data analysis included 144 images (it should have been 214 images). In this session, the subject still continued the task without being scanned for approximately 3 minutes so that our behavioural data analysis included all trials. All participants gave written informed consent for participation, and ethics approval for the study was granted by the Kochi University of Technology Ethics Board.

2.2. Procedure & Task

Participants were told they would view a series of scenarios which describe either a pro-social or anti-social behavior (e.g. "*Japanese students from University F were presented with the scenario of seeing racist material towards South Korean people on social media, and asked if they condoned this*", for full list of scenarios used see Supplementary materials) inside an fMRI scanner, and it was their task to estimate how typical Japanese and South Korean students answered a series of questions relating to how they would respond in said scenarios. They were asked to rate on a scale of 0%-100% in increments of 5 using a button box with three buttons. They used the index finger to increase the rating by 5%, the middle finger to reduce it by 5%, and the ring finger to give a final decision. All participants used their right hand to give responses. After they gave their rating, participants were presented with the "actual" rating given by Japanese or South Korean students, hereby referred to as *feedback* (see Figure 1 for visual of a complete trial). Although participants were led to believe that the feedback was real, in reality it was determined by a simple algorithm. Participants were exposed to 4 types of scenarios (2 [pro- versus anti- social] × 2 [Japan versus South Korea]), with a feedback trial that was higher or lower than the participant's first estimate. Our algorithm, computed via Matlab, ensured that feedback created roughly equal numbers of conditions across sessions, with a possible difference between participants' first ratings and feedback ranging from 5 to 30. The fMRI session consisted of a total of four runs, each consisting of 28 experimental trials plus 1 catch trial (where we presented feedback that coincided with participant’s first estimates). Participants were presented with the initial scenario for 3 seconds, with no limit when providing their ratings on how likely the group in question would partake in such scenario. Subjects response was highlighted for 1 second before feedback was presented for 2 seconds.

A total of 56 scenarios (plus 4 catch trials) were used in the fMRI experiment, and these scenarios were selected by a pilot study with an independent sample of n = 17 (mean age = 20.2, 9 males) from the Kochi University of Technology. In the pilot study, participants were asked to rate how likely a group of Japanese and South Korean students would respond to a total of 112 (56 Japanese and 56 South Korean) scenarios, as well as rate how positive / negative (valence rating) and relevant each scenario was on a scale of 1-7. Scenarios that presented extreme (ratings that fell outside of the bottom 7% and top 90%) ratings (how likely the target group in question responded) were discarded so as to reduce the effect of participants inevitably providing less extreme ratings in a subsequent second rating task, known as the regression-to-the-mean effect (RTM) which continually illustrates when repeated measures designs are used extreme values at the first measurement tend to approach the mean at the succeeding measurement (Galton, 1886; Yu and Chen., 2015). Scenarios were additionally matched for valence and relevance. This data was also used to generate extra scenarios that resembled and replicated the general theme of accepted scenarios, yielding a total of 28 positive Japanese scenarios, 28 negative Japanese scenarios, 28 positive South Korean scenarios, and 28 negative South Korean scenarios. Note that participants view the same positive scenarios for both the Japan and South Korea conditions, likewise for negative scenarios (i.e., “*Japanese students from University F were presented with the scenario of seeing racist material towards South Korea…”* versus “*South Korean students from University C were presented with the scenario of seeing racist material towards Japan…”* - the only aspect manipulated is the nationality of the students depicted in the scenario).

After the main fMRI session, participants were asked to re-rate all 112 scenarios they viewed in the scanner. This was to assess the effect of learning or update. In addition, they rated each of the 56 scenarios using a 7-point scale on how socially desirable the behavior depicted in each scenario was, excluding any nationality information (that of previous students completing the task and also the person depicted in the scenario) (1 = extremely socially undesirable, 4 = neither socially desirable nor undesirable, 7 = extremely socially desirable).

To assess their implicit attitudes toward Japan and South Korea, participants were asked to complete an Implicit Association Test (IAT) (Greenwald et al., 1998). The IAT included eight positive (e.g., Joy, Love, Wonderful) and eight negative words (e.g., Agony, Terrible, Nasty), all words were translated into Japanese. The Japan category included typical Japanese names (e.g., Shima, Nakata, Ono) whilst the South Korean category included typical Korean names (e.g., Han, Kim, Myong). All Japanese and South Korean names were matched on word length. Finally, their explicit attitudes toward Japan and South Korea were measured using a semantic differential scale. Participants rated each of Japan and South Korea on six bipolar dimensions using a 7-point scale; ugly-beautiful, bad-good, unpleasant-pleasant, honest-dishonest, foolish-wise, awful-nice and unfavorable-favorable. Finally, after completing a demographics questionnaire, to help ensure our experimental stimuli was efficient, participants were asked if they doubted anything during the experiment. They were debriefed, thanked and paid 2,000 yen for their participation.

[Insert Figure 1 here]

2.3. fMRI Data Acquisition

All fMRI data was acquired using a Siemens 3.0 Tesla Verio scanner with a 32 channel phased array head coil. For functional imaging, interleaved T2*- weighted gradient-echo echo-planar imaging (EPI) sequences were used to produce 40 contiguous 3.0-mm-thick trans-axial slices covering nearly the entire cerebrum (repetition time [TR] = 2,500 ms; echo time [TE] = 25 ms; flip angle [FA] = 90°; field of view [FOV] = 192 mm; 64 × 64 matrix; voxel dimensions = 3.0 × 3.0 × 3.0 mm). A high-resolution anatomical T1-weighted image (1 mm isotropic resolution) was also acquired for each participant.

2.4. fMRI Data Pre-processing

The fMRI data was analyzed using SPM12 (Welcome Department of Imaging Neuroscience) implemented in MATLAB (Math Works). Before data processing and statistical analysis, we discarded the first four volumes to allow for T1 equilibration. Head motion was corrected using the realignment program of SPM12. Following realignment, the volumes were normalized to MNI space using a transformation matrix obtained from the normalization of the first EPI image of each individual participant to the EPI template using an affine transformation (resliced to a voxel size of 2.0 × 2.0 × 2.0 mm). The normalized fMRI data were spatially smoothed with an isotropic Gaussian kernel of 8 mm (full-width at half-maximum).

2.5. fMRI Data Analysis

We used two general linear models (GLM) to analyze the fMRI data; one GLM was intended to identify brain regions correlated with the absolute differences between participant's estimate and feedback (hereby referred to as: *Absolute Gap*, see Figure 1B), and the other GLM was to explore brain regions correlated with the behavioral *Update* (difference between the first estimate and the second estimate, see figure 1B).

We used a parametric modulation analysis to investigate the relationship between trial-by-trial Absolute Gap scores and regional brain activity. We analyzed the fMRI data based on a 2 (Japan or South Korea) × 2 (favorable or unfavorable) design, yielding the four following conditions: 1) Japan-Favorable, 2) Japan-Unfavorable, 3) South Korea-Favorable, and 4) South Korea-Unfavorable, and data was first divided into four sets accordingly. The factor of favorable-unfavorable refers to the interaction between the valence of presented scenarios (positive or negative) and the feedback given in relation to participants first estimates (if this was better or worse than participants initial expectations), and whether this combination comes across as overall pro-social or antisocial. For example, a favorable trial would be depicted by higher feedback in a positive scenario (i.e., Japanese or South Korean students are *more* willing to act pro-socially than participants expected) or lower feedback in a negative scenario (i.e., Japanese or South Korean students are *less* willing to act antisocially than participants expected). Accordingly, the first model included: 1) each trial presentation (duration = total time from onset of initial scenario presentation to onset of feedback presentation), 2) Feedback presentation in Japanese favorable trials (duration = 2 sec), 3) Feedback presentation in Japanese favorable trials modulated by Absolute Gap, 4) Feedback presentation in Japanese unfavorable trials (duration = 2 sec), 5) Feedback presentation in Japanese unfavorable trials modulated by Absolute Gap, 6) Feedback presentation in South Korean favorable trials (duration = 2 sec), 7) Feedback presentation in South Korean favorable trials modulated by Absolute Gap, 8) Feedback presentation in South Korean unfavorable trials (duration = 2 sec), 9) Feedback presentation in South Korean unfavorable trials modulated by Absolute Gap, 10) Catch trial presentation (regressor of no interest) (duration = total time of catch trial from initial scenario presentation onset to the end of feedback presentation). This analysis yielded the four main contrast images (all conditions modulated by Absolute Gap) used for second level analysis. Other regressors that were of no interest, such as six motion parameters, the session effect, and high-pass filtering (128 sec) were also included.

The second GLM is similar to the first except we used the behavioral *Update* (the difference between the first vs. second estimates) as opposed to *Absolute Gap* (the difference between the first estimate vs. feedback) as a parametric regressor. Because the simple difference between the two estimates is susceptible to the RTM effect (Izuma & Adolphs, 2013; Yu & Chen, 2015), in order to remove the change between the first vs. second estimates which is explained by the RTM effect, we first run a linear regression analysis within each participant to estimate the RTM effect for each participant. The regression model used all 112 trials and included participant's first estimates as the only predictor variable, and Update as the dependent variable. All participants showed a negative beta value for first estimates (e.g., the higher the first estimate, the more likely participants decrease their estimate on the second rating task), and at a group level, it was significantly negative ($t(27) = -11.92$, $p < 0.001$), indicating the existence of the RTM effect. Within each participant, for each trial, we computed the Update scores predicted by the RTM effect and subtracted it from the actual Update scores (actual Update scores - Update scores predicted by the RTM effect). We then used the new controlled Update scores as parametric modulators in the second GLM. The same set up was utilized yielding the same contrast images to be used for second level analysis. For all fMRI analysis, a whole-brain statistical threshold was set at $p < 0.001$ voxel wise (uncorrected) and cluster $p < 0.05$ (FWE corrected for multiple comparisons).

In addition to these two main GLMs, we also ran three additional GLMs (see Supplementary Materials for the full details and results of these GLMs); one addressed the effect of "general favorability" of feedback (i.e. if feedback indicated more people are willing to engage in a socially desirable behavior or less people are willing to engage in an anti-social behavior, *regardless of participant's expectations*). The second GLM incorporated both Absolute Gap and Update in a single GLM, and the third incorporated Update and Favorability in a single GLM to assess the interaction of Update x Favorability on brain activity.

2.6. Behavioral Data Analysis

For the IAT, a score for each participant was calculated using the D-score algorithm developed by Greenwald, Nosek, and Banaji (2003). Positive IAT D-scores indicate more positive implicit evaluation of Japan relative to South Korea. Semantic differential scores for each participant were computed by averaging the six bipolar scales separately for Japan and South Korea.

To calculate the effect of feedback on the extent participants updated their second estimates, two multiple regressions (one for Japanese Trials, and one for South Korean trials) were run to analyze behavioral data. Both included predictor variables: 1) First Estimates, 2) Gap (feedback - first estimate, not absolute value), 3) Favorability (dummy coded as favorable = 1 and unfavorable = 0), and 4) Gap × Favorability. All predictors were centered by subtracting the mean value from each score to evade multicollinearity. The dependent variable was Update (second estimate – first estimate).

We additionally ran a similar analysis to assess the effect of "general favorability" of trials as mentioned above (see Supplementary Materials for the full details and results of this analysis).

Due to our stimuli incorporating scenarios that do versus don't involve the other-group in some form (i.e. "*... If you saw racist material towards Japanese people on social media, would you feel positive about it?* ", versus, "*... Do you believe it is acceptable that when intoxicated at a party people sometimes vandalize property?* "), we conducted analysis to compare any potential confounds from this. We divided the data into scenarios that did involve the other-group (n=15), and scenarios that didn't (n=13). The same analysis as described above for both Japanese and South Korean trials was applied within each set of data, for full details, see Supplementary Materials.

3. RESULTS

3.1. Behavioral Results

3.1.1. Attitudes towards Japan versus South Korea

We first found that, not surprisingly, Japanese participants' explicit evaluations of Japan were significantly more positive than those of South Korea: ($t(27) = 7.95$, $p < 0.001$, Cohen's d = 1.97) (Figure 2a). We further demonstrate that explicit evaluations of Japan are significantly positive (by examining how different the mean score was from the midpoint of the scale: [$t(27) = 11.55$, $p < 0.001$]), and that those of South Korean were significantly negative ($t(27) = -2.11$, $p = 0.04$). Additionally, IAT scores were significantly positive ($t(27) = 4.14$, $p < 0.001$, Cohen's d = 0.80) (Figure 2b), indicating more positive implicit evaluations of Japan relative to South Korea. No significant correlation was observed for implicit evaluations and explicit evaluations (Japanese minus South Korean mean scores) ($r = 0.10$, $p = 0.62$), and no significant correlation was observed for explicit evaluations between Japan versus South Korea ($r = 0.16$, $p = 0.41$).

[Insert Figure 2 here]

3.1.2. Effect of Gap on Update

Our multiple regression analyzes utilizing *Update* as the dependent variable revealed a significant effect of *Gap* (feedback - first estimate) for Japanese ($t(27) = 10.97$ $p < 0.001$) and for South Korean trials ($t(27) = 11.0$ $p < 0.001$), meaning that participants updated their scores *more* from the first to the second rating the larger the gap was between their first rating and the feedback they were presented with. The effect of Favorability was not significant for both Japan and South Korea trials (Table 1). However, we observed a significant interaction effect of Gap and Favorability (whether the interaction between the scenario and feedback is overall Favorable or Unfavorable) for Japanese trials ($t(27) = 3.25$, $p = 0.003$) meaning that participants updated their scores significantly more in response to favorable feedback compared to unfavorable feedback. The same interaction effect for the South Korea condition was in the same direction, but didn't reach the significance ($t(27)= 1.54$, $p = 0.13$). There was no significant difference in the Gap × Favorability interaction effect between the Japanese and South Korean conditions ($p = 0.30$). Accordingly, although our results showed significantly more positive implicit and explicit evaluations of Japan compared to South Korea (Figure 2, also see Table 1), contrary to our prediction, the level of favorability bias is no different between ingroup and outgroup. Thus, our behavioral results showed that participants tended to update their scores more if the feedback allows them to see other people (regardless of nationality) more favorably. Of final note, it should be stated that no significant difference at group level was observed for any of the Japanese and South Korean predictors (First Estimate $p = 0.23$; Gap $p = 0.68$; Favorability $p = 0.43$; see Table 1).

Table 1. Behavioral regression model statistics demonstrating beta and p values for all predictor variable.

Predictor Variable	Mean Standardized Beta Value	Standard Deviation	*p* value
Japanese			
First Estimate	-7.40	3.84	<0.001**
Gap	7.45	3.60	<0.001**
Favorability	0.67	1.81	0.060
Gap × Favorability	2.06	3.36	0.003**
South Korean			
First Estimate	-8.11	3.72	<0.001**
Gap	7.17	3.45	<0.001**
Favorability	0.28	1.86	0.043*
Gap × Favorability	1.35	4.62	0.134

All values are based on a multiple regression analysis within each participant. P values are based on group level one-sample t-tests. Japanese mean $R^2 = 0.46$, Japanese mean Adjusted $R^2 = 0.42$. South Korean mean $R^2 = 0.44$, South Korean mean Adjusted $R^2 = 0.40$. * $p < 0.05$, ** $p < 0.01$

3.1.3. Correlation of Explicit Attitudes and Favorability Bias Index

Although we didn't observe a significant difference in favorability bias between ingroup and outgroup, we observe significant across-subject correlations between explicit evaluations and favorability bias for both Japan ($r = 0.33$, $p = 0.04$) and South Korea ($r = 0.53$, $p = 0.002$), respectively (Figure 3). These results are, at least partially, consistent with our prediction and indicate that the strength of favorability bias depends on individuals' attitudes toward a group; the higher the explicit evaluation of Japan or South Korea, the more participants updated their belief about members of each group when the feedback is in a favorable direction compared to an unfavorable direction.

The Japanese vs. South Korean favorability bias indices were significantly correlated with each other ($r = 0.61$, $p < 0.001$), while as stated above, the corresponding explicit evaluations were not significantly correlated with each other ($r = 0.16$, $p = 0.41$), indicating that there exists individual differences in viewing other people favorably in general.

Thus, our behavioral results indicate that participants update their ratings more when feedback is in a *favorable* direction as opposed to an *unfavorable* direction, and this effect is seemly consistent across nationalities (Table 1). Nonetheless, individual differences in the tendency to update ratings in a favorable direction compared to an unfavorable direction (i.e., favorability bias) were correlated with participants' explicit evaluations for each of the Japan and South Korea conditions (Figure 3).

Finally of note, to further examine any bias elicited by participants first estimates, we ran a within-subject correlational analysis to check if participants' first estimates are correlated with Absolute Gap. But, we found no significant correlation for both Japanese (p= 0.32) or South Korean trials (p= 0.38).

[Insert Figure 3 here]

3.2. fMRI Results

3.2.1. Imaging results depicting the effect of Gap

In order to first broadly depict regions related to the conflict between one's initial rating in relation to feedback, we used *Absolute Gap* (absolute value) as a parametric modulator. We investigated the effect of Absolute Gap regardless of condition (i.e., by combining all of the four conditions [Japanese-Favorable, Japanese-Unfavorable, South Korean-Favorable, and South Korean-Unfavorable]). Here, we found that pMFC (specifically the dmPFC and left supplementary motor area; SMA), lateral superior temporal gyrus (STG), and posterior cingulate cortex (PCC) activity is positively correlated with Absolute Gap (see Table 2 & Figure 4A, B, & C). These regions are largely consistent with areas previously implicated in social conflict (the difference between one's and others' opinions) in a social conformity paradigm (Izuma and Adolphs, 2013; Klucharev et al., 2009; Wu et al., 2016). For full information of the overlap between the current studies activation map and that of Izuma and Adolph (2013), see Supplementary Results. In our main ROI of the dmPFC (x = -8, y = 24, z = 66), the effect of Gap was significantly positive in all conditions excluding Japanese Favorable, which was marginally insignificant (Japanese Favorable $p = 0.08$, all remaining $ps < 0.001$; Figure 4C).

Furthermore, examination of brain regions *negatively* correlated with Absolute Gap revealed significant activation within the ventral striatum (specifically nucleus accumbens, see both Table 2 for full list of regions activated and Figure 5A & B for associated contrast image), also consistent with previous studies. For results of regions correlated with Absolute Gap for each condition separately (Japanese Favorable, Japanese Unfavorable, South Korean Favorable, South Korean Unfavorable), see Supplementary Table 4.

Table 2. Brain regions correlated with Absolute Gap

Location	BA	MNI coordinate x	y	z	Z	Cluster size
Areas *positively* correlated with Absolute Gap						
dmPFC	8	-8	24	66	5.16	1996
left supplementary motor area (SMA)	8	-6	22	58	5.12	
left superior frontal gyrus (SFG)	9	-12	46	46	4.87	
Right superior temporal gyrus (STG)	20	44	16	-36	4.84	327
Left STG	30	-42	20	-30	5.07	1569
left inferior frontal gyrus (IFG)	47	-44	32	-6	4.89	
left insula	47	-40	22	-8	4.87	
Posterior cingulate cortex (PCC)	23	-6	-50	28	4.80	1076
Areas *negatively* correlated with Absolute Gap						
Right postcentral gyrus	40	56	-40	50	6.18	2321
right supramarginal gyrus	40	46	-36	40	5.60	
right angular gyrus	40	40	-48	54	5.07	
Left postcentral gyrus	40	-48	-36	44	5.82	2431
left angular gyrus	40	-54	-40	54	5.79	
Right middle frontal gyrus (MFG)	8	26	16	56	5.98	557
Left MFG	46	-38	34	26	5.55	931
Right ventral striatum	25	12	10	-10	5.42	1051
right IFG	44	52	12	24	4.98	

BA, Brodmann area. Statistics are based on a set threshold of height $p < 0.001$ (uncorrected), and cluster $p < 0.05$ (FWE). Areas in grey italics represent significant peak (FWE) sub-clusters/different regions within larger clusters.

[Insert Figure 4 here]

Interestingly, exploration of the contrast image depicting activation for Unfavorable trials modulated by Absolute Gap compared to Favorable trials modulated by Absolute Gap (Unfavorable > Favorable) also revealed that a different cluster within the dmPFC (x = 6, y = 38, z = 48, k = 238), left inferior frontal gyrus (IFG, x = -48, y = 18, z = 22, k = 1137) and right middle frontal gyrus (MFG, x = 40, = 8, z = 58, k = 607) more sensitive to Absolute Gap in an *Unfavorable* direction compared to a Favorable direction (see Figure 6A & B). As shown in Figure 6C, the dmPFC tracked Absolute Gap in an Unfavorable direction, while it was insensitive to Absolute Gap in a Favorable direction. In contrast, no clusters survived the threshold in place when examining brain regions correlated with Absolute Gap for Favorable trials compared to Unfavorable trials (for full list of results, see Supplementary Results, Supplementary Table 5).

[Insert Figure 5 here]

We additionally explored several brain-behavior correlations. Although our behavioral results revealed robust individual differences in favorability bias, there was no significant correlation between the behavioral favorability bias and neural favorability bias (i.e., Unfavorable-Absolute Gap vs. Favorable-Absolute Gap) in the dmPFC (or any additional ROIs reported in Table 2) for both the Japan ($r = 0.21$, $p = 0.29$) and South Korea ($r = -0.00$, $p = 0.98$) conditions.

Thus, while our behavioral data showed that participants' updated their estimates more when the feedback was in a *favorable* direction, our fMRI data actually indicated that the cluster within the dmPFC (x = 6, y = 38, z = 48; Figure 6A) was more sensitive to the discrepancy between one's initial estimate and the feedback when the feedback was in an *unfavorable* direction.

[Insert Figure 6 here]

3.2.2. Imaging Results depicting the effect of Update

In order to further assess whether any brain regions are related to the actual *change* of participant’s ratings (Update), the same parametric modulation analysis was conducted using Update (controlled for RTM) as the parametric modulator, instead of Absolute Gap. No significant clusters survived the threshold, and although alluded to in some previous research regarding the pMFC and attitude change, no significant activation in these regions was observed via the same contrast image combining all conditions modulated by Update.

4. DISCUSSION

The aim of the study was to test two competing hypotheses regarding pMFC activity, those being; if the pMFC encodes the conflict between reality and a desirable outcome, or if the pMFC plays a role in belief updating. This was assessed by employing a cognitive bias to specifically disentangle the level of conflict from the level of belief updating, whilst assessing pMFC sensitivity respectively. Accordingly, our behavioral data indicates participants' are more likely to update their beliefs in the direction of *favorable* new information (especially in the Japan condition), whilst our fMRI data indicates that the dmPFC is more sensitive to *unfavorable* new information (Figure 6A), and this effect was consistent across Japanese and South Korean conditions. In contrast, no brain region was significantly related to behavioral update. Thus, the findings support the conflict hypothesis rather than the update hypothesis, indicative that sensitivity in the pMFC (at least within the dmPFC; Figure 6A) is related to the conflict between ideal scenarios versus reality.

Activation of the dmPFC in Izuma and Adolph (2013) tracked the discrepancy between one's own preference and its social ideal as defined by balance theory (Heider, 1946). In the current study we see a matching activation map to that of Izuma and Adolph (2013) across all combined conditions modulated by Absolute Gap (basically the degree of conflict in each trial, hereby referred to as such for the purpose of the discussion) (Figure 4). However, the same neural activation in regards to solely the updating of beliefs based on new information was not observed. Henceforth, it would seem likely that brain activity demonstrated in the current experiment is liable representative of the conflict of information presented, rather than any associated updating of beliefs. Nonetheless, it should be specified that the analysis is based on the onset of feedback presentation, not when participants give their second estimates, where any additional neural mechanisms (potentially the dmPFC) related to the update of belief may be more apparent. Although we focused on brain activations during the feedback processing in the first rating task just like a majority of previous social conformity studies, this might explain why under the current paradigm, no significant neural activity regarding the updating of beliefs was seen. It is interesting and important to see in future research whether the dmPFC, or other brain regions, tracks the degree of behavioral adjustments (update) similar to the ones implemented in the current study during the second rating task.

A key result from this study was that the dmPFC, left IFG, and right MFC were more sensitive to the degree of conflict in unfavorable compared to favorable trials. This tallies with Holroyd and Cole (2002), who highlight the pMFC's involvement with the focus on consequence predication in terms of action monitoring, specifically, when the outcome of a given task is *worse than expected*. An effect also relevant to this paradigm is the “False Consensus Effect” (Ross et al., 1977), the notion that people tend to believe more people share their attitudes/world view than actually do. Interestingly, Welborn and Lieberman (2018) found when examining the neural effects of consensus bias, pMFC (specifically the medial prefrontal cortex and ventral medial prefrontal cortex: mPFC, vmPFC) activity was positively associated with observed consensus bias only when information given to participants as feedback (similar to this study) was of a challenging/disconfirmatory nature, as opposed to confirming previous beliefs. Thus, our work appears to replicate a specific sensitivity of goal-driven conflict within the pMFC, also fitting nicely with a recent review regarding the motivational characteristics of cognitive consistency, that being we strive more for specifically *favored* outcomes rather than consistent ones alone (Kruglanski et al., 2018).

Although the present study demonstrated that these regions were more sensitive to unfavorable information, it was favorable information that was more successfully updated in the second rating task. The contrast between our fMRI and behavioral data on the surface resembles the general effect of cognitive dissonance (discomfort evoked by the discrepancy between attitudes, beliefs, and behavior) (Festinger, 1962), a form of conflict in its simplest form. That being, participants seemingly exhibit more negative emotion from the unfavorable feedback (indicated by increased sensitivity in the aforementioned ROIs), yet do not update it as efficiently. This allies with previous research which also posits the pMFC (Harmon-Jones et al., 2008) as being a central neural correlate of cognitive dissonance, particularly in the dmPFC (Izuma et al., 2010) and dACC (Izuma et al., 2010; Van Veen et al., 2009; Izuma & Murayama, *in press*). However, it should be said that in more typical examples of cognitive dissonance, participants often resolve this by amending behavior and/or attitudes accordingly, whereas in the current study participants seem to resolve this conflict by not (or to a lesser extent) updating their belief according to unfavorable information (further discussion on the lack of memory update is extended in the next paragraph). One important distinction to first make here is that participants' also have an additional conflict of being "correct", since there is a factually correct answer in this experimental paradigm, whereas classic cognitive dissonance studies tend to revolve around preference (which participants can freely change). This avoids any extra level of divergence the current participants' may have underwent (resolving dissonance vs. being correct), which could possibly have added to the lack of update observed in the current experiment.

Relatedly, and in somewhat contrast to the current study, Hughes et al., (2017) found participants were more likely to update their impressions regarding negative information during an impression formation task about out-group members, but not in-group members. This was associated with less engagement in the dACC, temporoparietal junction, insula, and precuneus when processing negative information about the in-group, but importantly not the out-group. The asymmetry of participants impression update and neural response between in versus out-group members suggests that these neural structures are important for updating one’s impression, especially when new information fits with individual's pre-existing notions (e.g., in-group positive behavior and out-group negative behavior). Though this study is similar in many ways to the current experiment, there are several key differences. First relates to the point above regarding the re-assessment of subjective (opinion) versus objective (facts) information, which is an important distinction between Hughes et al., and the current study. Second, it should also be noted that though we do measure subjective impressions (explicit attitudes) of the out-group (and in-group) as they do in Hughes et al., (2017), because this was only measured at a single timepoint in the current experiment, it isn’t possible to compare any possible update/change of this after participants received feedback. Moreover, it’s also relevant to highlight that the participants who produced lower explicit attitudes towards the out-group did tend to update more unfavorable information, allying with Hughes et al., (2017) findings.

In order to continue to elucidate the role of the dmPFC, it is increasingly important to assess the effect of memory. In an apparent contrast to our results, previous research would suggest that more conflicting or shocking information is more likely to be remembered (Berntsen, 2002; Kensinger, 2007). This might suggest that unfavorable information was not updated due to participants' active inhibition of the effect of unfavorable information on update during the second estimation task. Alternatively (but not necessarily mutually exclusive), what may be apparent is inefficient encoding of the feedback during the first estimation task. Our data demonstrates that activity in the left IFG, and the dmPFC was more sensitive to Gap in unfavorable trials compared to favorable trials (Figure 6), and these two regions have been implicated in response inhibition (Floden and Stuss, 2006; Verfaellie and Heilman, 1987). Historically, increased activation in the right (as opposed to the left) IFG has been associated with increased inhibitory control of responses (e.g. De Zubicaray et al., 2000; Garavan et al., 1999; Konishi et al., 1999), but there is some suggestion that the left IFG also plays a central role in response inhibition. Specifically, Swick et al., (2008) found patients with left IFG legions had higher error rates than controls in both conditions (easy vs. hard) of a Go/NoGo task, being further impaired in the hard condition when more inhibitory control was required. Future research should examine more extensively neural activities during the second rating task and the relationship regarding the valence of social information and subsequent memory processes (e.g., whether unfavorable feedback is better remembered) to tease apart the two possibilities (increased inhibition vs. decreased encoding).

Further ROIs we found from the fMRI data include areas of the striatum (nucleus accumbens specifically) which were negatively correlated with the degree of conflict in each trial (Figure 5). This supplements previous research that also demonstrates when participants' opinions differ from that of others, whilst the pMFC is activated, the striatum is deactivated (Campbell-Meiklejohn et al., 2010; Izuma and Adolphs, 2013; Klucharev et al., 2009). Welborn and Lieberman (2018) infer their similar finding in terms of the gratifying value of information. This seems a tenable explanation, with additional links made toward reinforcement learning surrounding conformity by Klucharev et al., (2009). Alternativly, it seems an important distinction that our dmPFC (Figure 4) and ventral striatum clusters encode Absolute Gap across all trials (positively: dmPFC, or negatively: ventral striatum) in a relatively objective manner (i.e., unaffected by favorability of information), suggesting these regions are related to general learning mechanisms. One the other hand, the dmPFC cluster that encodes Absolute Gap specifically for Unfavorable compared to Favorable trials (Figure 6) seems to be influenced by a top down emotional process so that in addition to the objective difference (Absolute Gap), the activity is modulated by what participants *hope* the reality to be. Thus, our ventral striatum activation may represent the processing of information more objectively (rather than subjectively being influenced by the valuation of information). This relates nicely to a recent fMRI study by Pine et al., (2018), which specifically highlights the ventral striatum's involvement in the learning of factual knowledge.

Our results also demonstrate increased sensitivity for the degree of conflict within the PCC and lateral STG. The PCC has been implicated in tracking the cognitive imbalance between own preferences versus others, as well as being correlated with subsequent preference changes in Izuma and Adolph's (2013). Furthermore, work by Falk et al., (2014) show the PCC is more sensitive to social exclusion in participants who also subsequently change their actions to suit peers (in this case, increase the level of risk in their driving more around peers as opposed to alone). Although our data doesn't demonstrate an association with the behavioral update, it seems consistent that this region plays a role in the recognition of social conflict. Not only has this been established in terms of social conflict (see also Seehausen et al., 2014), neuroimaging studies have also shown the PCC to be sensitive in monitoring nonsocial prediction errors and conflict in general (Christoffels, Formisano, & Schiller, 2007; Kadosh, Kadosh, Henik, & Linden, 2008). The STG has some similar implications in the monitoring of social conflict (Christoffels et al., 2007). For example, Premkumar et al., (2012) report the right STG to be more active during the viewing of social rejection as opposed to neutral scenes, and Seehausen et al., (2014) found the STG to be more active in an empathy-experiment where participnats felt misunderstood (in comparison to understood)- both implicating a potential role in the discrimination of desirable versus undesirable outcomes.

Behaviorally, participants demonstrated a favorability bias in general. We display a correlation between positive evaluations to Japan *or* South Korea and the extent participants update their beliefs based on more favorable information. More broadly put, participants increasingly revise their belief based on new information to see people more positive for previously more liked social groups, supplementing the previously discussed work of Izuma and Adolph (2013). As participants overall possessed positive explicit evaluations of Japan, the data coincides with our behavioral hypothesis that more beliefs are updated regarding favorable information. However, although our participants explicit and implicit evaluations were on average significantly less positive for South Korea, participants did still elicit a favorability bias at the group level for South Korea also, updating their beliefs more so for favorable trials here too.

Our initial behavioral hypothesis stated that any favorability effect would be less pronounced for South Korea owing to less positive attitudes in general. This outcome was forecast to arise due to the effect of confirmation bias, seeing participants update information that more aligns with their previous attitudes (more positive towards Japan versus less positive towards South Korea). An initial consideration here, then, is that the results are more consistent with the “good-news-bad-news-effect” (Eil and Rao, 2011). This is the concept that information and its corresponding valence are not updated and processed in an equal, linear manner. Positive information (good news) tends to revise according to previous experience and is more efficiently updated, whereas the updating of negative information (bad news) is not, being more noisy and less likely to be updated into current beliefs. Broadly applied to the current findings, this would suggest that updating favorable compared to unfavorable information takes place in a more efficient and uniform manner, regardless of any pre-existing views and thus the social group applied to. This has been supported by work on optimism bias (Sharot et al., 2011), demonstrating participants’ are more likely to update their belief based on more positive information about the future compared to negative information. This positivity bias is theorized to arise as a protection for general mental well-being (Garrett et al., 2018; Sharot et al., 2011).

It should also be noticed that the explicit evaluations towards South Korea displayed large across-participant variability, with many participants having close-to-neutral attitudes (meaning they didn't feel particularly positive *or* negative towards South Korea). But to reiterate, the participants who *did* have extremely negative explicit evaluation’s towards South Korea did tend to update their beliefs more in response to *unfavorable* feedback. Speculatively, since we only measured explicit attitudes at a single time point, these results might suggest that more moderate attitudes are increasingly amendable upon receiving information, more easily disconfirming any preexisting *weaker* stereotypes. This, in comparison to more extreme attitudes in which the information may be updated more asymmetrically (as presented by Sunstein et al., 2016), further facilitating attitude polarization, additionally coincides with research that demonstrates increased dogmatic-intolerance in relation to attitude extremity (van Prooijen and Krouwel, 2017).

Future research may wish to select a more exclusively hostile and defined in/outgroup paradigm in order to further extract any additional effects of attitude extremity, and the associated neural correlates/behavioral update. For example, it may be interesting to examine a potential ceiling (or cross-over) effect of the good-news-bad-news model in terms of extreme attitudes- at what point is bad news about a disliked outgroup no longer perceived as “bad”, but instead information that only affirms ones previous distain? What’s more, if the pMFC is sensitive to social conflict as we showed, this should in theory then be less robust for negative information regarding disliked outgroups for people with extremely negative attitudes due to lesser conflict between ones social outlook versus reality. Finally, although we found similar neural correlates of Absolute Gap (Figures 4 & 5) between the present study with Japanese participants and our previous study with American participants (Izuma & Adolphs, 2013), it is important to systematically and directly test cultural differences in social information processing in future research, as previous studies indicated cultural differences in social conformity (Bond and Smith, 1996; Korn et al., 2014) and cognitive dissonance (Kitayama et al., 2004 but see also Chen and Risen, 2010; Izuma and Murayama, 2013).

CONCLUSION

In sum, the current experiment demonstrated two key points, i) activity in the dmPFC was representative of socially conflicting information, specifically the conflict between ideal outcomes versus less ideal realities, and not the corresponding belief update based on new information. ii) participants updated their beliefs based on more favorable information, of which related to more positive evaluations of the social group in question. Future research should aim to further disentangle the role of the dmPFC in social conflict processing, attempting to apply experimental paradigms to specifically isolate potentially independent neural correlates related to the actual update of participants beliefs based on new information received. What can be taken from the current study overall is an increased understanding of the role played by the dmPFC in social information processing, of which ultimately helps us to understand how decisions about social interactions are made, providing a more solid foundation for social attitude amendment and interventions.

AKNOWLEDGEMENTS

This work was supported by the JSPS London JBUK Japan Award (to K.I.), JSPS KAKENHI Grant Number 15K12777 and 17H00891 (to K.N.) and University Grant for Research Facility and Project, Kochi University of Technology (to R.A. and K.N.).

REFERENCES

Berntsen, D., 2002. Tunnel memories for autobiographical events: Central details are remembered more frequently from shocking than from happy experiences. Mem. Cognit. 30, 1010–1020.

Bond, M.H., Smith, P.B., 1996. CROSS-CULTURAL SOCIAL AND ORGANIZATIONAL PSYCHOLOGY, Annu. Rev. Psychol.

Campbell-Meiklejohn, D.K., Bach, D.R., Roepstorff, A., Dolan, R.J., Frith, C.D., 2010. How the opinion of others affects our valuation of objects. Curr. Biol. 20, 1165–1170.

Chen, M.K., Risen, J.L., 2010. How choice affects and reflects preferences: Revisiting the free-choice paradigm. J. Pers. Soc. Psychol. 99, 573–594. https://doi.org/10.1037/a0020217

Chen, J., Wu, Y., Tong, G., Guan, X., Zhou, X., 2012. ERP correlates of social conformity in a line judgment task. BMC Neurosci. 13, 43.

Christoffels, I.K., Formisano, E., Schiller, N.O., 2007. Neural correlates of verbal feedback processing: an fMRI study employing overt speech. Hum. Brain Mapp. 28, 868–879.

De Zubicaray, G.I., Andrew, C., Zelaya, F.O., Williams, S.C.R., Dumanoir, C., 2000. Motor response suppression and the prepotent tendency to respond: a parametric fMRI study. Neuropsychologia 38, 1280–1291.

Eil, D., Rao, J.M., 2011. The good news-bad news effect: asymmetric processing of objective information about yourself. Am. Econ. J. Microeconomics 3, 114–138.

Falk, E.B., Cascio, C.N., O’Donnell, M.B., Carp, J., Tinney Jr, F.J., Bingham, C.R., Shope, J.T., Ouimet, M.C., Pradhan, A.K., Simons-Morton, B.G., 2014. Neural responses to exclusion predict susceptibility to social influence. J. Adolesc. Heal. 54, S22–S31.

Festinger, L., 1962. A theory of cognitive dissonance. Stanford university press.

Floden, D., Stuss, D.T., 2006. Inhibitory control is slowed in patients with right superior medial frontal damage. J. Cogn. Neurosci. 18, 1843–1849.

Galton, F., 1886. Regression towards mediocrity in hereditary stature. J. Anthropol. Inst. Gt. Britain Irel. 15, 246–263.

Garavan, H., Ross, T.J., Stein, E.A., 1999. Right hemispheric dominance of inhibitory control: an event-related functional MRI study. Proc. Natl. Acad. Sci. 96, 8301–8306.

Garrett, N., González-Garzón, A., Foulkes, L., Levita, L., Sharot, T., 2018. Updating Beliefs Under Perceived Threat.

Greenwald, A.G., McGhee, D.E., Schwartz, J.L.K., 1998. Measuring individual differences in implicit cognition: the implicit association test. J. Pers. Soc. Psychol. 74, 1464.

Greenwald, A.G., Nosek, B.A., Banaji, M.R., 2003. Understanding and using the implicit association test: I. An improved scoring algorithm. J. Pers. Soc. Psychol. 85, 197.

Hackel, L.M., Zaki, J., Van Bavel, J.J., 2017. Social identity shapes social valuation: evidence from prosocial behavior and vicarious reward. Soc. Cogn. Affect. Neurosci. 12, 1219–1228. https://doi.org/10.1093/scan/nsx045

Harmon- Jones, E., Gerdjikov, T., Harmon- Jones, C., 2008. The effect of induced compliance on relative left frontal cortical activity: A test of the action- based model of dissonance. Eur. J. Soc. Psychol. 38, 35–45.

Heider, F., 1946. Attitudes and cognitive organization. J. Psychol. 21, 107–112.

Holroyd, C.B., Coles, M.G.H., 2002. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol. Rev. 109, 679.

Huang, Y., Kendrick, K.M., Yu, R., 2014. Social conflicts elicit an N400-like component. Neuropsychologia 65, 211–220.

Huang, Y., Zhen, S., Yu, R., 2019. Distinct neural patterns underlying ingroup and outgroup conformity. Proc. Natl. Acad. Sci. U. S. A. 116, 4758–4759. https://doi.org/10.1073/pnas.1819421116

Hughes, B.L., Zaki, J., Ambady, N., 2017. Motivation alters impression formation and related neural systems. Soc. Cogn. Affect. Neurosci. 12, 49–60. https://doi.org/10.1093/scan/nsw147

Izuma, K., 2013. The neural basis of social influence and attitude change. Curr. Opin. Neurobiol. 23, 456–462.

Izuma, K., Adolphs, R., 2013. Social manipulation of preference in the human brain. Neuron 78, 563–573.

Izuma, K., Aoki, R., Shibata, K., Nakahara, K., 2019. Neural signals in amygdala predict implicit prejudice toward an ethnic outgroup. Neuroimage 189, 341–352. https://doi.org/10.1016/j.neuroimage.2019.01.019

Izuma, K., Matsumoto, M., Murayama, K., Samejima, K., Sadato, N., Matsumoto, K., 2010. Neural correlates of cognitive dissonance and choice-induced preference change. Proc. Natl. Acad. Sci. 201011879.

Izuma, K., Murayama, K., 2013. Choice-Induced Preference Change in the Free-Choice Paradigm: A Critical Methodological Review. Front. Psychol. 4, 41. https://doi.org/10.3389/fpsyg.2013.00041

Izuma, K., & Murayama, K. (2019). The neural basis of cognitive dissonance. In E. Harmon-Jones (Ed.), Cognitive dissonance: Reexamining a Pivotal Theory in Psychology (2nd edition). Washington, DC: American Psychological Association.

Kensinger, E.A., 2007. Negative emotion enhances memory accuracy: Behavioral and neuroimaging evidence. Curr. Dir. Psychol. Sci. 16, 213–218.

Kitayama, S., Snibbe, A.C., Markus, H.R., Suzuki, T., 2004. Self and Dissonance in Two Cultures.

Kim, B.-R., Liss, A., Rao, M., Singer, Z., Compton, R.J., 2012. Social deviance activates the brain’s error-monitoring system. Cogn. Affect. Behav. Neurosci. 12, 65–73.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., Fernández, G., 2009. Reinforcement learning signal predicts social conformity. Neuron 61, 140–151.

Knobloch-Westerwick, S., Mothes, C., Johnson, B.K., Westerwick, A., Donsbach, W., 2015. Political online information searching in Germany and the United States: Confirmation bias, source credibility, and attitude impacts. J. Commun. 65, 489–511.

Konishi, S., Nakajima, K., Uchida, I., Kikyo, H., Kameyama, M., Miyashita, Y., 1999. Common inhibitory mechanism in human inferior prefrontal cortex revealed by event-related functional MRI. Brain 122, 981–991.

Korn, C.W., Fan, Y., Zhang, K., Wang, C., Han, S., Heekeren, H.R., 2014. Cultural influences on social feedback processing of character traits. Front. Hum. Neurosci. 8, 192. https://doi.org/10.3389/fnhum.2014.00192

Kruglanski, A.W., Jasko, K., Milyavsky, M., Chernikova, M., Webber, D., Pierro, A., Santo, D., Kruglanski, A.W., Jasko, K., Milyavsky, M., Chernikova, M., Kruglanski, A.W., Jasko, K., Milyavsky, M., Chernikova, M., Webber, D., 2018. An International Journal for the Advancement of Psychological Theory Cognitive Consistency Theory in Social Psychology : A Paradigm Reconsidered. Psychol. Inq. 29, 45–59. https://doi.org/10.1080/1047840X.2018.1480619

Lee, C.-S., 1985. Japan and Korea. The Political Dimension., Pacific Affairs. Hoover press. https://doi.org/10.2307/2758349

Lin, L.C., Qu, Y., Telzer, E.H., 2018. Intergroup social influence on emotion processing in the brain. Proc. Natl. Acad. Sci. 115, 10630–10635. https://doi.org/10.1073/pnas.1802111115

Lord, C.G., Ross, L., Lepper, M.R., 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. J. Pers. Soc. Psychol. 37, 2098.

Mansouri, F.A., Egner, T., Buckley, M.J., 2017. Monitoring demands for executive control: shared functions between human and nonhuman primates. Trends Neurosci. 40, 15–27.

Molenberghs, P., Louis, W.R., 2018. Insights From fMRI Studies Into Ingroup Bias. Front. Psychol. 9, 1868. https://doi.org/10.3389/fpsyg.2018.01868

Pine, A., Sadeh, N., Ben-Yakov, A., Dudai, Y., Mendelsohn, A., 2018. Knowledge acquisition is governed by striatal prediction errors. Nat. Commun. 9, 1673. https://doi.org/10.1038/s41467-018-03992-5

Premkumar, P., Ettinger, U., Inchley- Mort, S., Sumich, A., Williams, S.C.R., Kuipers, E., Kumari, V., 2012. Neural processing of social rejection: the role of schizotypal personality traits. Hum. Brain Mapp. 33, 695–706.

Ross, L., Greene, D., House, P., 1977. The "false consensus effect": An egocentric bias in social perception and attribution processes. J. Exp. Soc. Psychol. 13, 279–301. https://doi.org/10.1016/0022-1031(77)90049-X

Sambrook, T.D., Goslin, J., 2015. A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. Psychol. Bull. 141, 213.

Schnuerch, R., Gibbons, H., 2015. Social proof in the human brain: Electrophysiological signatures of agreement and disagreement with the majority. Psychophysiology 52, 1328–1342.

Schnuerch, R., Trautmann-Lengsfeld, S.A., Bertram, M., Gibbons, H., 2014. Neural sensitivity to social deviance predicts attentive processing of peer-group judgment. Soc. Neurosci. 9, 650–660.

Seehausen, M., Kazzer, P., Bajbouj, M., Heekeren, H.R., Jacobs, A.M., Klann-Delius, G., Menninghaus, W., Prehn, K., 2014. Talking about social conflict in the MRI scanner: neural correlates of being empathized with. Neuroimage 84, 951–961.

Sharot, T., Korn, C.W., Dolan, R.J., 2011. How unrealistic optimism is maintained in the face of reality. Nat. Neurosci. 14, 1475.

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

Shestakova, A., Rieskamp, J., Tugin, S., Ossadtchi, A., Krutitskaya, J., Klucharev, V., 2012. Electrophysiological precursors of social conformity. Soc. Cogn. Affect. Neurosci. 8, 756–763.

Sunstein, C.R., Bobadilla-Suarez, S., Lazzaro, S.C., Sharot, T., 2016. How people update beliefs about climate change: Good news and bad news. Cornell L. Rev. 102, 1431.

Swick, D., Ashley, V., Turken, U., 2008. Left inferior frontal gyrus is critical for response inhibition. BMC Neurosci. 9, 102.

Tajfel, H., 1982. SOCIAL PSYCHOLOGY OF INTERGROUP RELATIONS.

Tajfel, H., 2010. Social identity and intergroup relations. Cambridge University Press.

van Prooijen, J.-W., Krouwel, A.P.M., 2017. Extreme political beliefs predict dogmatic intolerance. Soc. Psychol. Personal. Sci. 8, 292–300.

Van Veen, V., Krug, M.K., Schooler, J.W., Carter, C.S., 2009. Neural activity predicts attitude change in cognitive dissonance. Nat. Neurosci. 12, 1469.

Verfaellie, M., Heilman, K.M., 1987. Response preparation and response inhibition after lesions of the medial frontal lobe. Arch. Neurol. 44, 1265–1271.

Welborn, B.L., Lieberman, M.D., 2018. Neuropsychologia Disconfirmation modulates the neural correlates of the false consensus e ff ect : A parametric modulation approach. Neuropsychologia 121, 1–10. https://doi.org/10.1016/j.neuropsychologia.2018.09.018

Wu, H., Luo, Y., Feng, C., 2016. Neural signatures of social conformity: A coordinate-based activation likelihood estimation meta-analysis of functional brain imaging studies.

Neurosci. Biobehav. Rev. 71, 101–111.

Yu, R., Chen, L., 2015. The need to control for regression to the mean in social psychology studies. Front. Psychol. 5, 1574.

Figure 1

Click here to download high resolution image

Figure 2

Click here to download high resolution image

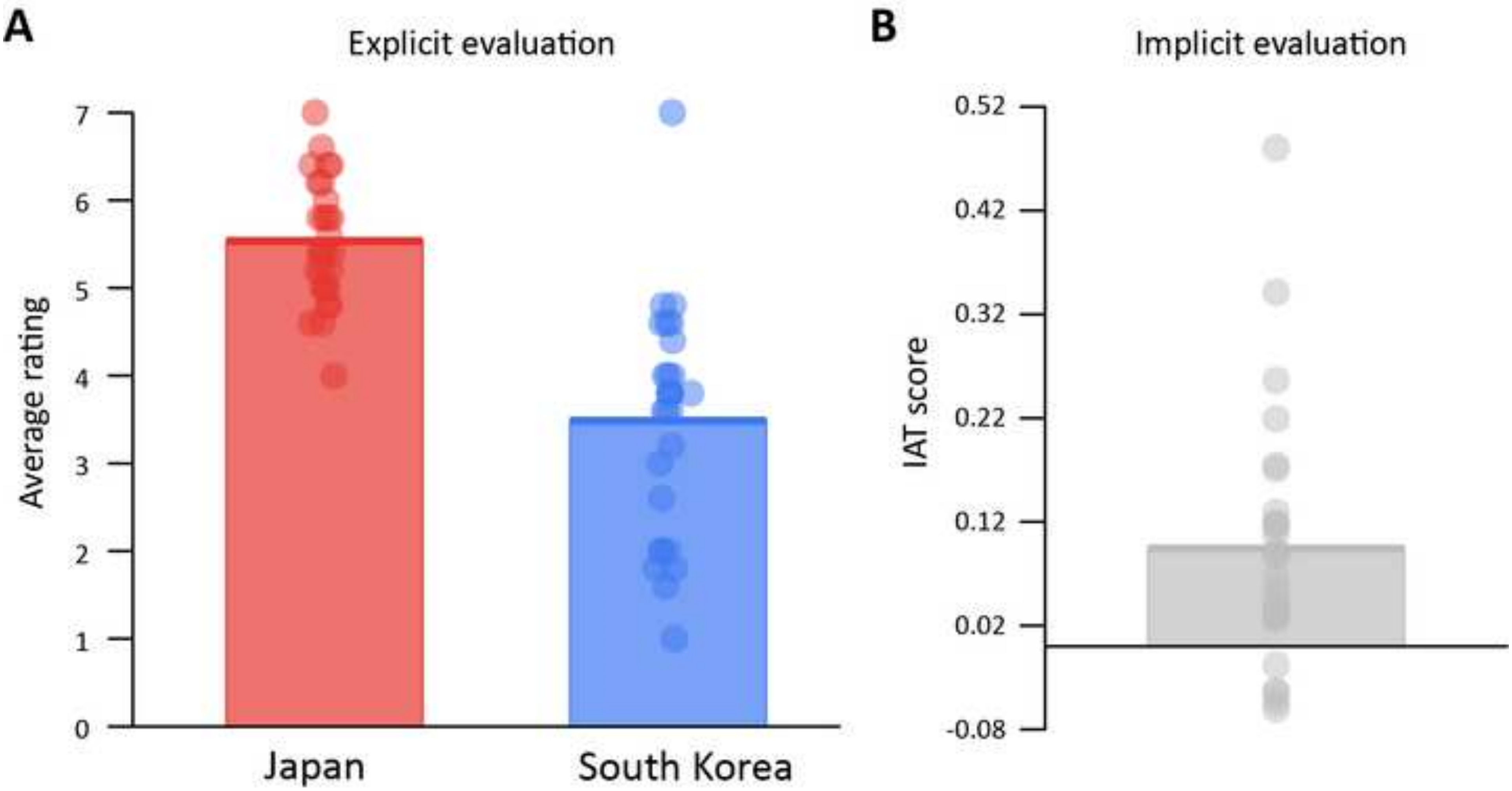

Figure 3

Click here to download high resolution image

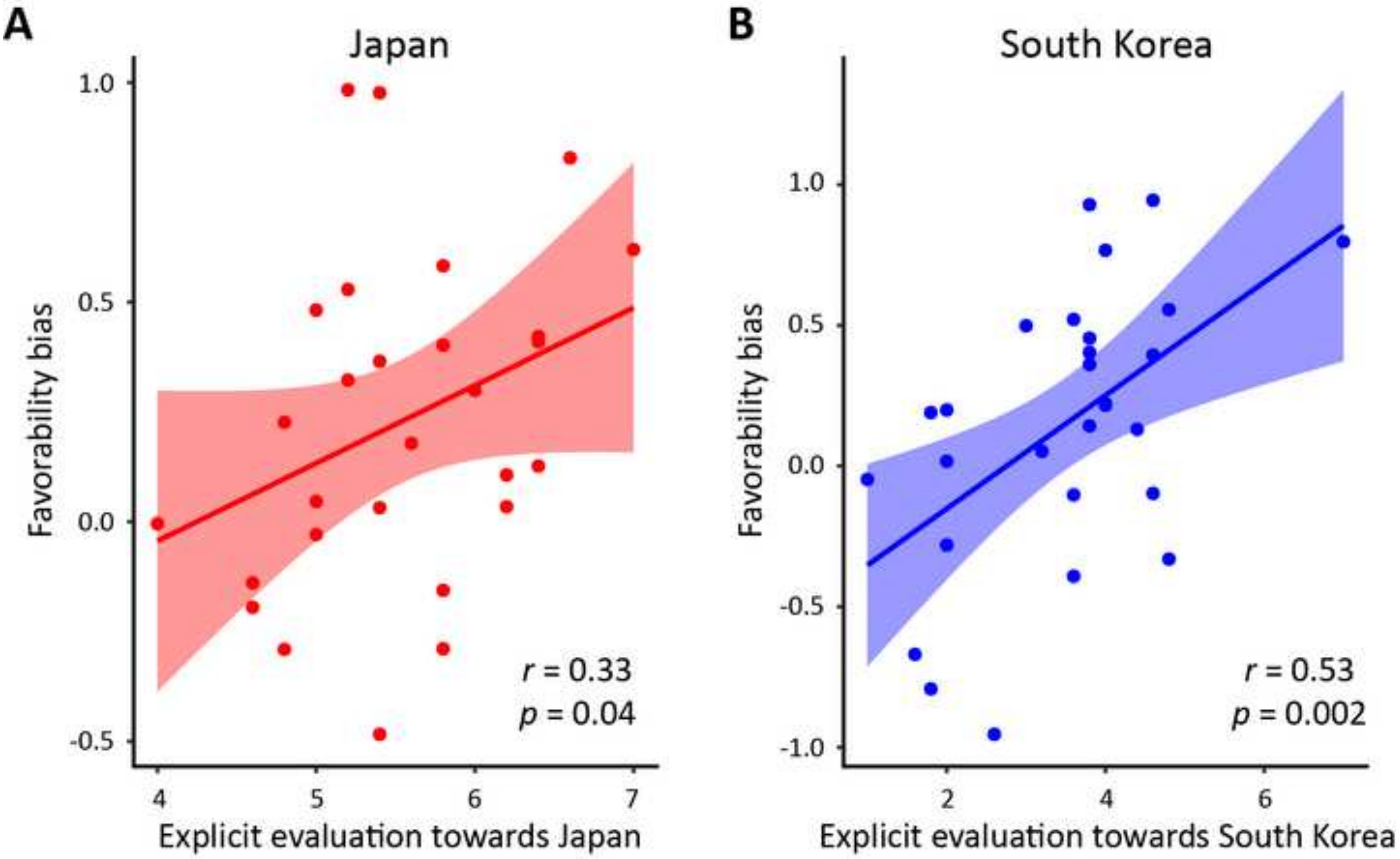

Figure 4

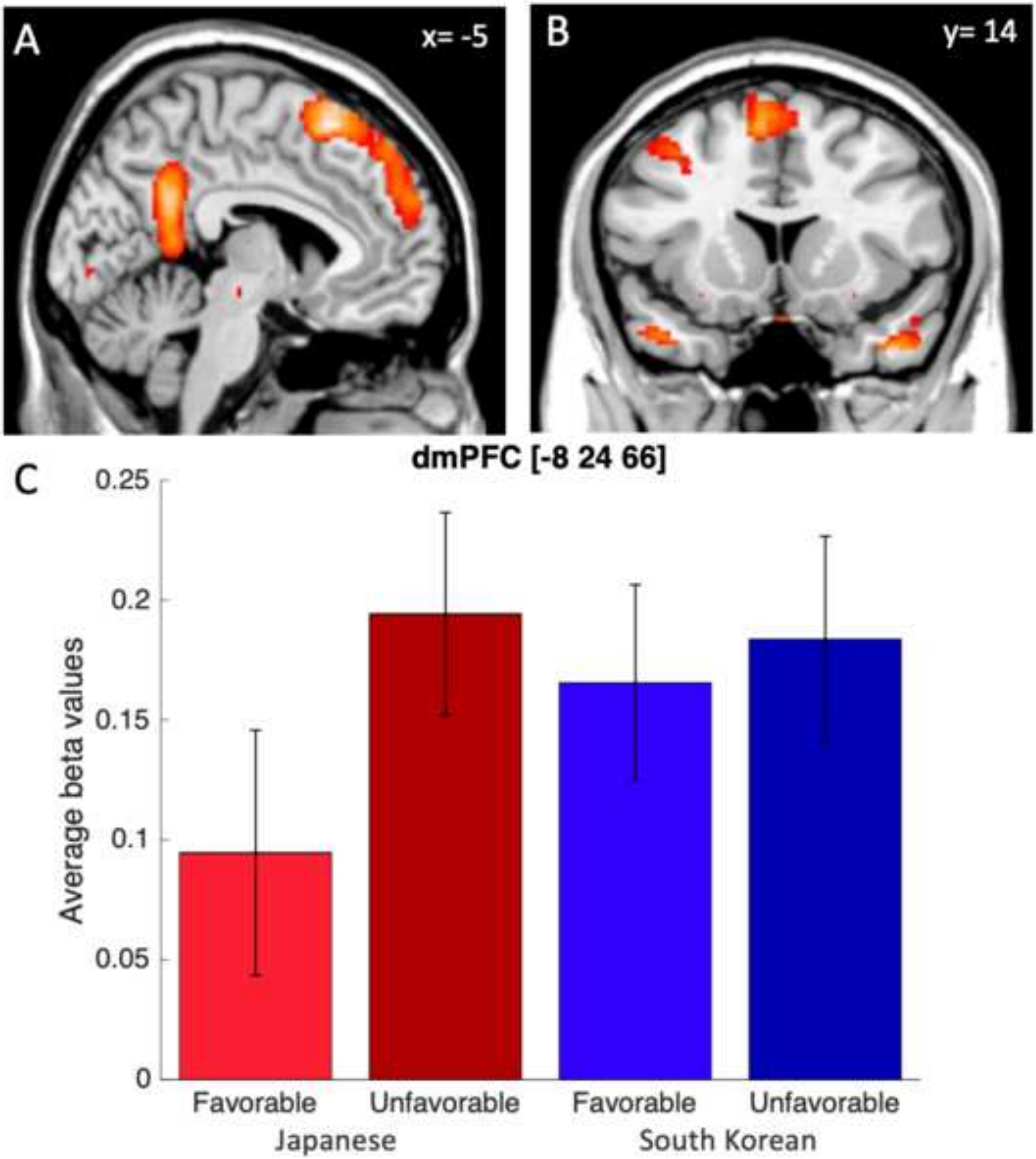
A
x= -5
B
y= 14
dmPFC [-8 24 66]
C
0.25
0.2
0.15
0.1
0.05
0
Average beta values
Favorable
Unfavorable
Favorable
Unfavorable
Japanese
South Korean

Figure 5

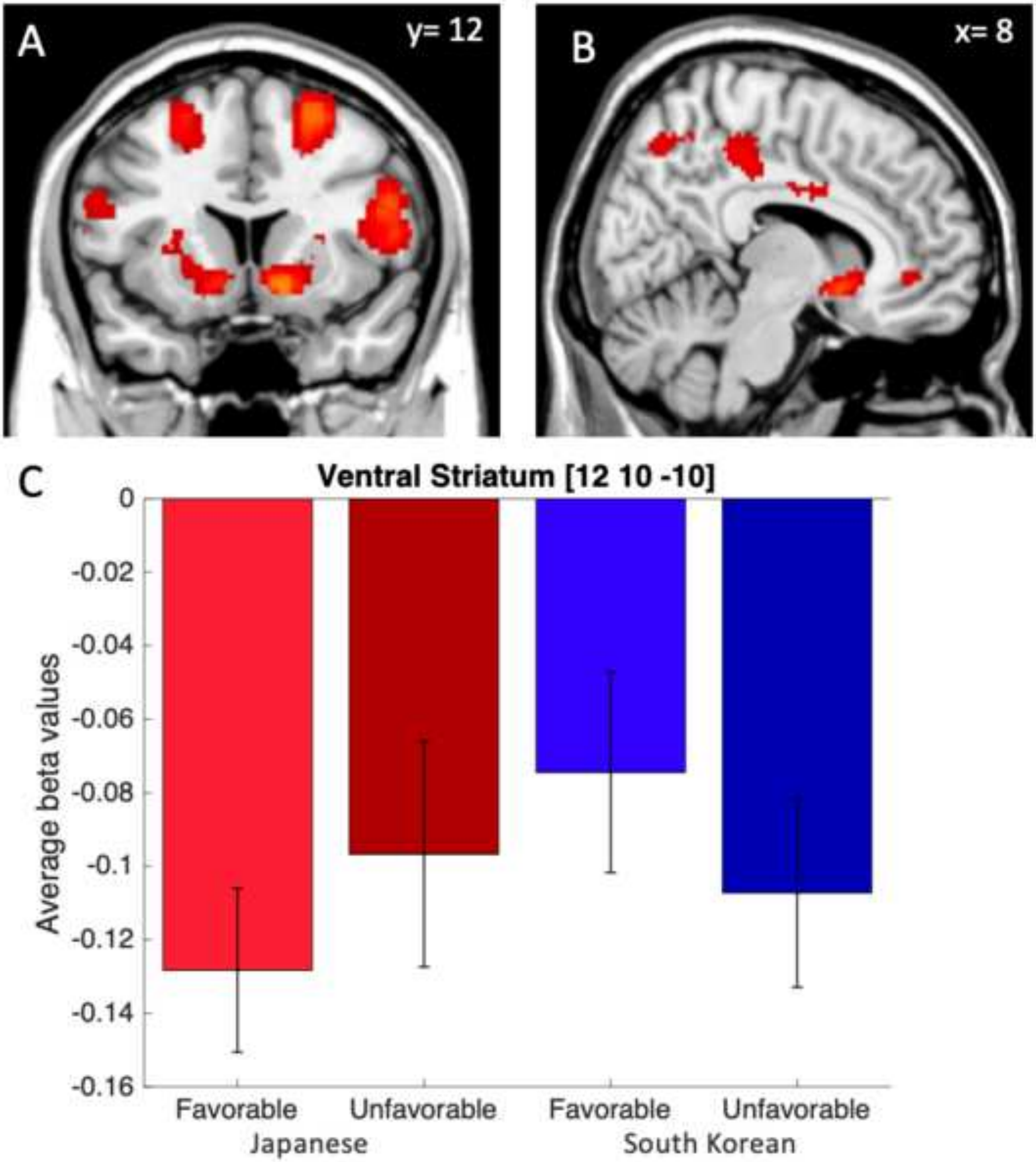
A
y= 12
B
x= 8
C
Ventral Striatum [12 10 -10]
Average beta values
0
-0.02
-0.04
-0.06
-0.08
-0.1
-0.12
-0.14
-0.16
Favorable
Unfavorable
Favorable
Unfavorable
Japanese
South Korean

Figure 6

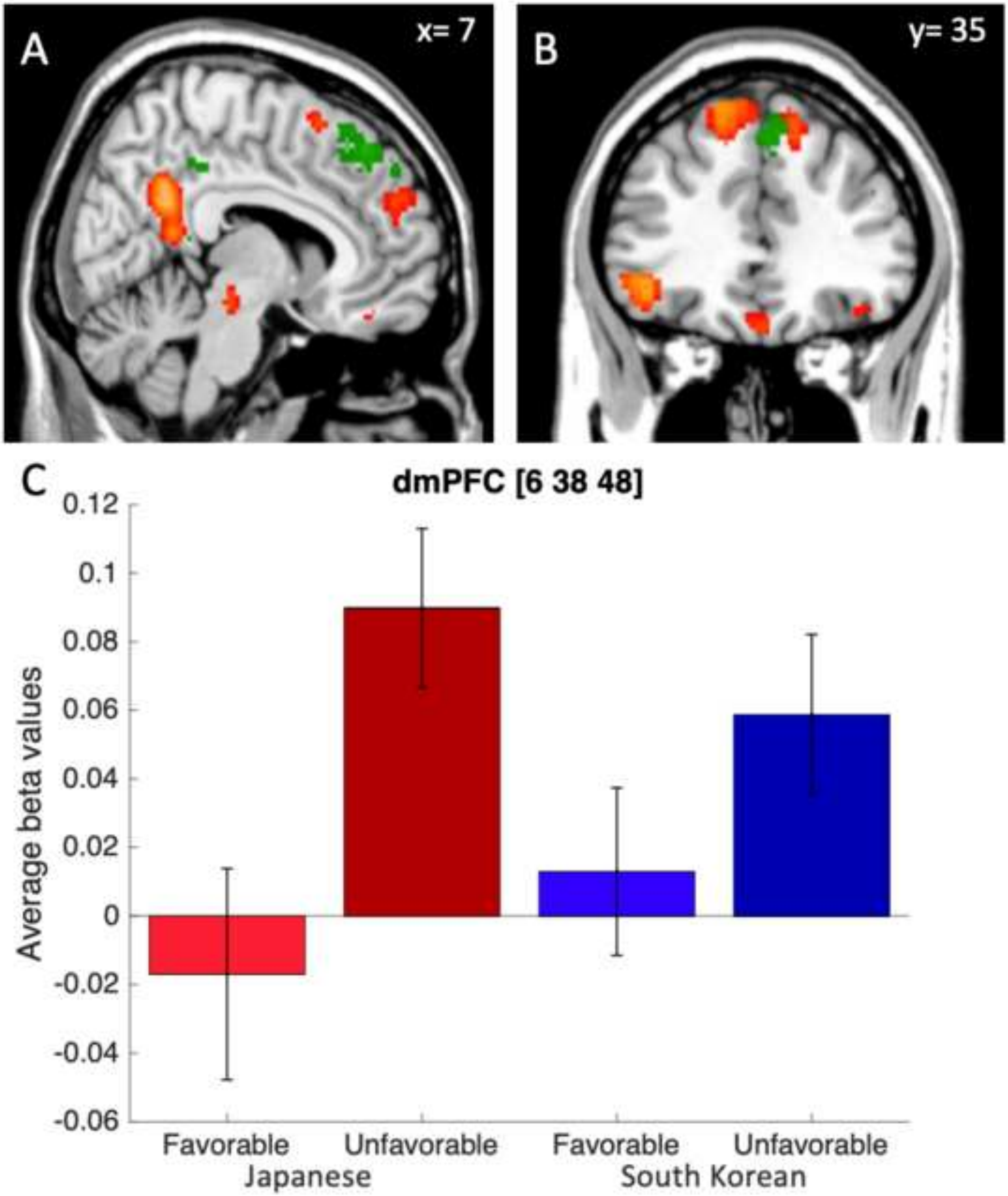
A
x= 7
B
y= 35
C
dmPFC [6 38 48]
Average beta values
0.12
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
Favorable
Unfavorable
Favorable
Unfavorable
Japanese
South Korean

Figure legends

Figure 1. A) Example of a complete South Korean trial (scenario, question / first rating, feedback) utilized for fMRI stimuli, as seen by participants inside the scanner. Each trial started with a scenario presentation (description of a pro- or anti-social behavior) for 3 seconds, after which participants' were asked to give their first estimation of how likely the person in question (Japanese vs. South Korean student) rated they would partake in said behavior (in which they had no time limit). After, the estimate was highlighted in yellow for 1 second followed by feedback presentation (the "true value") for 2 seconds. **B**). Visual representation of Absolute Gap and Update scores. **C**). Example of 4 scenario types depicted via a pro-social scenario. Feedback was reversed in order to create the same conditions for anti-social scenarios.

Figure 2. A) Bars represent mean explicit evaluations (semantic differentials). Higher numbers indicate more positive evaluation. **B**) Bar represents mean IAT D-score. Positive scores indicate more positive implicit evaluation of Japan relative to South Korea. Circles denote individual data points.

Figure 3. Scatter plot demonstrating positive correlation between participants' explicit evaluations of Japan (**A**) and South Korea (**B**), and favorability bias (i.e. the extent participants update their beliefs in favorable trials compared to unfavorable trials). Shaded areas represent 95% confidence intervals.

Figure 4. A) Sagittal slice (x = -5) demonstrating brain regions positively correlated with Absolute Gap. **B**) Coronal slice (y = 14) demonstrating brain regions positively correlated with Absolute Gap**. C**) Bars represent average beta values across all conditions within key significant cluster in the dmPFC, error bars denote SEM. All betas were extracted via a 4mm sphere from the peak activation identified by the contrast image depicting all trials modulated by Absolute Gap.

Figure 5. A) Coronal slice (y = 12) demonstrating brain regions negatively correlated with Absolute Gap. **B**) Sagittal slice (x = 8) demonstrating brain regions negatively correlated with Absolute Gap. **C**) Bars represent average beta values across all conditions within key significant cluster in the ventral striatum. All betas were extracted via a 4mm sphere from the peak activation identified by the contrast image depicting all trials modulated by Absolute Gap, and error bars denote SEM.

Figure 6. A) Sagittal slice (x = 7) demonstrating brain regions correlated with Absolute Gap (all of the four conditions combined; shown in orange), as well as brain activity for unfavorable compared to favorable trials modulated by Absolute Gap (shown in green). This contrast partially replicates Figure 4A (activation shown in orange) from a slightly different slice perspective in order to demonstrate the independent nature of the dmPFC sensitivity specifically for unfavorable trials (green) compared to across all trials (orange). **B**) Coronal slice (y = 35) demonstrating brain regions correlated with Absolute Gap (all of the four conditions combined; shown in orange), as well as brain regions significantly more strongly correlated with Absolute Gap in unfavorable trials compared to favorable trials (shown in green). **C**) Bars represent average beta values across all conditions within key significant cluster in the dmPFC (x = 6 y = 38 z = 48). All betas were extracted via a 4mm sphere from

the peak activation identified by the contrast image depicting unfavorable compared to favorable trials modulated by Absolute Gap. All error bars denote SEM.