Running Title: Neural correlates of the FCE

Neural Correlates of the False Consensus Effect:
Evidence for Motivated Projection and Regulatory Restraint

B. Locke Welborn[1]

Benjamin C. Gunter[2]

I. Stephanie Vezich[2]

Matthew D. Lieberman[2]

[1]University of California, Santa Barbara

[2]University of California, Los Angeles

Correspondence should be addressed to:

Matthew D. Lieberman
Department of Psychology
4611 Franz Hall, UCLA
Los Angeles, CA 90095-1563

Phone: 310.825.9257
Email: lieber@ucla.edu

Abstract

The False Consensus Effect (FCE), the tendency to project our attitudes and opinions on to others, is a pervasive bias in social reasoning with a range of ramifications for individuals and society. Research in social psychology has suggested that numerous factors (anchoring and adjustment, accessibility, motivated projection, etc.) may contribute to the FCE. In the present study, we examine the neural correlates of the FCE and provide evidence that motivated projection plays a significant role. Activity in reward regions (VMPFC and bilateral NAcc) during consensus estimation was positively associated with bias, while activity in RVLPFC (implicated in emotion regulation) was inversely associated with bias. Activity in reward and regulatory regions accounted for half of the total variation in consensus bias across subjects (R^2=.503). This research complements models of the FCE in social psychology, providing a glimpse into the neural mechanisms underlying this important phenomenon.

INTRODUCTION

Adaptation to the pressures and pitfalls of a dynamic social environment demands acute sensitivity to the attitudes, perspectives, and opinions of others. Whether official pollsters or ordinary social thinkers, we expend a great deal of effort to understand how others feel about the issues of the day. Nevertheless, empirical research shows that our understanding of others' attitudes is consistently biased by the positions we hold ourselves, a phenomenon known as the false consensus effect (FCE, or 'consensus bias'; Ross, Greene, & House, 1977; Marks & Miller, 1987). Moreover, this consensus bias has proven remarkably recalcitrant, persisting stubbornly when challenged by social feedback, sometimes even in the face of unanimous disagreement (Krueger & Clement, 1994). Given the importance of understanding others' attitudes, why should our own attitudes exert such a profound impact on our perceptions of social reality, and what mechanisms support this bias?

One prominent theory contends that consensus bias is a consequence of motivated projection – in short, we misperceive others' attitudes because we want to think of ourselves as being in the majority, holding views that are normatively 'right' (see Crano, 1983; Sherman, Presson, & Chassin, 1984; Morrison & Matthes, 2011). If the projection of our own attitudes onto others is an instance of motivated projection, we might expect that consensus bias would be associated with neural correlates of social reward. Indeed, social approval has been found to activate neural structures involved in reward learning (Izuma et al., 2008; Simon, Becker, Mothes-Lasch et al., 2014) such as the nucleus accumbens (NAcc) and ventromedial prefrontal cortex (VMPFC). Sharing our own attitudes with others has also been associated with reward (participants were willing to forego monetary payment to self-disclose), with corresponding

activity in NAcc (Tamir & Mitchell, 2012). If motivated projection contributes significantly to the FCE by enhancing feelings of social approval or as a prelude to social sharing, we might therefore expect between-subjects differences in the FCE to covary with activity in these reward regions.

Conversely, in order to accurately estimate the attitudes of others, regulatory mechanisms may be necessary in order to overcome the affective lure of our own antecedent opinions. That is, our own attitudes may serve as an evaluative anchor when we consider the attitudes of others, and regulatory mechanisms may help us to detach from this starting point and assess the attitudes of others more objectively (Tversky & Kahneman, 1974; Tamir & Mitchell, 2010). Functional neuroimaging studies have consistently implicated the right and left ventrolateral prefrontal cortex (RVLPFC and LVLPFC) in emotion regulation (see meta-analysis by Kohn, Eickoff, Scheller, et al., 2014). Both of these regions have also been invoked when individuals must detach from their own perspective (Cohen, Berkman, & Lieberman, 2012; Hartwright, Apperly, Hansen, 2015). If regulatory mechanisms are required in order to inhibit the prejudicial pull of one's antecedent attitudes and to accept the possibility that one's own position may not be predominant, then activity in RVLPFC and LVLPFC may be inversely associated with exhibited consensus bias.

Guided by the social psychological literature on the false consensus effect, we sought to test the putative processes of motivated projection and regulatory restraint during consensus estimation in a functional neuroimaging study. We therefore interrogated hemodynamic response using fMRI while participants estimated the attitudes of the ordinary member of a comparison population (other UCLA undergraduates) on contemporary social, personal, and political issues.

Laboratory studies of the FCE typically ask for estimates of consensus in the absence of contextual information, but we were also interested in the effects of participants having some information that might be relevant to making the consensus judgment. To this end, we varied the information available about the attitudes of other UCLA students on a trial-by-trial basis, providing participants with false feedback concerning their peers. On "Confirmation" trials, participants were led to believe that another individual held an attitudinal position comparable to their own, a manipulation that we hoped would reaffirm participants' (biased) intuition that their attitudes were normative or commonplace amongst their peers. In contrast, on "Disconfirmation" trials participants were informed that another individual held an attitude discrepant with their own, a manipulation we believed might encourage participants to restrain (insofar as possible) the tendency toward motivated projection. For the last trial type, "No Information" trials, participants made consensus estimates without additional feedback.

If motivated projection and regulatory restraint are important contributors to consensus bias, we anticipated that reward regions such as NAcc and VMPFC would drive consensus bias, while regulatory regions such as LVLPFC and RVLPFC would attenuate bias. In addition, the role of these regions and their putative psychological processes in consensus bias may interact with informational context. During Confirmation trials, social reward processes may exacerbate consensus bias uninhibited by contradictory feedback, whereas Disconfirmation trials may push participants towards more critical interrogation of their attitudes and increase the likelihood of successful regulation.

METHODS

Participants

Twenty-nine participants (17 female) were recruited by email and Internet solicitations from the psychology research subject pool at UCLA. All participants had been enrolled as undergraduate students at UCLA for at least two quarters, and none had taken an introductory course in social psychology (in order to preclude familiarity with the false consensus effect). Participants were judged ineligible if they did not differ from our estimate of the mean UCLA undergraduate attitude on a sufficient number of items. All participants were compensated $40 for their contribution to this research or received course credit. Participants provided written informed consent approved by the UCLA Institutional Review Board. One participant's data are not included in these analyses due to partial acquisition failure (final n=28).

Attitude Item Selection

Attitude items were selected from a larger set of 155 social, political and personal issues (e.g. abortion rights, gay marriage, daily flossing, making out on a first date) that had previously been tested with an online sample of 178 UCLA undergraduates. Participants in this online sample indicated their attitudes towards each issue using a numeric scale ranging from 0 to 100 in integer increments (with anchors 0 – Complete Opposition, 25 – Moderate Opposition, 50 – Neutrality, 75 – Moderate Support, and 100 – Complete Support). These responses provided a reasonable estimate of the mean UCLA undergraduate attitude on each of the 155 issues, and these values were used to determine error of estimation for the scanner task described below.

Prior to scanning, prospective participants in the present study indicated their own attitudes on each of the 155 issues, and were eligible to participate only if their responses differed from our estimate of the UCLA undergraduate population mean by at least 15 points on at least 90 items. If participants did not differ in their attitudes from the group mean for the items used, it would not possible to disambiguate projection from accurate consensus estimation on a

trial-by-trial level. As this was a major objective of the study, we felt it was necessary to impose such an inclusion criterion in order to provide a sufficient number of viable trials for the scanner task. The idiosyncrasies of participants' attitudes on the stimulus issues resulted in the selection of a unique set of attitude items for each individual, on each of which they differed from the UCLA undergraduate mean by at least 15 points. These items were randomly and equivalently divided amongst the Confirmation, Disconfirmation, and No Information conditions. Across participants, this procedure resulted in an average of 99 trials total, or 33 per Consensus Estimation condition.

Consensus Estimation Task:

While undergoing functional magnetic resonance imaging (fMRI), participants estimated the attitude of the ordinary UCLA student on each of the ideographically-selected attitude items (see above). During the 'No Information' condition, participants were simply asked to provide their best possible estimate of the attitude that an ordinary UCLA student would have on the given issue. In order to do this, they used an on-screen scale identical to that used during item selection (as described above) except that the values represented the attitude that the ordinary UCLA student would have, rather than the participant's own attitude.

In the 'Confirmation' and 'Disconfirmation' conditions, participants were provided with on-screen information ostensibly reflecting the attitudes of other UCLA undergraduates. Participants were told that, on each trial, the attitude of a different UCLA student from our larger Internet sample would be presented, and that they could use (or disregard) this information in making their consensus estimates. While this sample actually existed, and was used to determine the true norms for each attitude item as described above, participants actually received false information designed to either Confirm or Disconfirm the presupposition that their own attitudes

would be representative of the UCLA undergraduate population as a whole. In the Confirmation condition participants were provided with an attitude that differed from their own by at most 5 points (in either direction). As all attitude items were pre-selected so that participants attitudes were at least 15 points different from the mean, this ensured that the sample attitudes presented in the Confirmation were closer to the participant's own attitude than to the mean UCLA undergraduate attitude. In the Disconfirmation condition participants were provided with a sample attitude that differed from the actual mean UCLA undergraduate attitude by at most 5 points (in either direction), so that this sample attitude was invariably closer to the actual mean than to the participant's own attitude. In both Confirmation and Disconfirmation conditions, deviations from the participant's own attitude and the mean UCLA undergraduate attitude were selected from a uniform random distribution so as to ensure that the presented attitude fell within the desired range.

On each trial (see Figure 1), the sample information (ostensibly reflecting the attitude of a single UCLA undergraduate) was presented numerically above the appropriate portion of the scale, with a line denoting the precise location corresponding to the other student's attitude. After the scale (and if applicable, sample information) had appeared on-screen, participants had 10 seconds within which to make their response. Trials were not explicitly separated into feedback and response phases, and sample information remained on-screen until participants had confirmed their response. Trial presentation was self-paced, with a jitter duration commencing immediately after participants' responses were registered. Inter-trial jitter was selected from an exponential random distribution with a range of 4-9s and a mean value of 5 seconds.

Non-social color-judgment trials were also included as a basic perceptual-motor control condition. On these trials, participants were asked to judge the color of an on-screen square that

varied continuously from completely red to completely blue. Participants were instructed to treat the mid-point value of '50' as indicating that the square appeared to them completely purple, and neither more blue nor more red in hue. If the square appeared more red than blue, participants were to select values greater than 50, with 100 indicated that they perceived the square to be completely red. If the square appeared more blue than red, participants were to select values less than 50 with 0 indicated that the square completely blue. Participants were instructed explicitly to provide *their own* judgment regarding the color of the square, and to ignore how others might perceive it. Thirty control trials were included in the task for each participant, intermixed with consensus estimation trials.

Trial order was pseudo-randomized such that no condition repeated more than twice sequentially and conditions were represented equally over two functional runs.

Post-scanning measures:

After completion of the Consensus Estimation task, participants viewed each attitude item again and indicated a) their confidence in the accuracy of their consensus estimation, and b) the subjective importance of their attitude on the issue. For both judgments, participants used a 100-point integer scale with anchors at 0 – Not at all confident (important), 25 – A little confident (important), 50 – Moderately confident (important), 75 – Very confident (important), and 100 – Extremely confident (important).

fMRI data acquisition

All imaging data was acquired using a 3.0-Tesla Siemens Trio scanner at the Ahmanson-Lovelace Brain Mapping Center at UCLA. Across 2 functional runs, approximately 650 T2*-weighted echo-planar images were acquired during completion of experimental tasks described above (slice thickness=3mm, gap=1mm, 36 slices, TR=2000ms, TE=25ms, flip angle=90°,

matrix=64x64, field of view=200mm). An oblique slice angle was used in order to minimize signal drop-out in ventral medial portions of the brain. In addition, a T2-weighted, matched-bandwith anatomical scan was acquired for each participant (TR=5000ms, TE=34ms, flip angle=90°, matrix=128x128; otherwise identical to EPIs). Lastly, we acquired a T1-weighted magnetically-prepared rapid acquisition gradient echo anatomical image (slice thickness=1mm, 176 slices, TR=2530ms, TE=3.31ms, flip angle=7°, matrix=256x256, field of view=256mm).

fMRI Data Preprocessing and Analysis

Preprocessing and Region of Interest (ROI) definition:

Functional data were analyzed using SPM8 (Wellcome Department of Cognitive Neurology, London, UK). Within each functional run, image volumes were corrected for slice acquisition timing, realigned to correct for head motion, segmented by tissue type, and normalized into standard MNI stereotactic space (resampled at 3x3x3mm). Finally, images were smoothed with an 8mm Gaussian kernel, FWHM.

Given our specific hypotheses regarding the role of reward and regulatory regions in shaping expression of consensus bias, all principal analyses were conducted on *a priori* regions-of-interest (ROIs). Reward regions were selected from a comprehensive meta-analysis of the literature on subjective valuation conducted by Bartra and colleagues (Bartra, McGuire, & Kable, 2013). Six millimeter spherical ROIs were defined based upon statistical meta-analytic peaks from their analysis of the decision phase of rewarding trials, during which participants selected between various choice alternatives on the basis of their subjective value (see Table 3 in Batra, McGuire, & Kable, 2013 for details). We felt that this particular conceptualization best matched the mechanism of motivated projection hypothesized to underlie the false consensus effect. This procedure yielded a VMPFC ROI centered at MNI:-2,40,-8 and left and right nucleus

accumbens (NAcc) ROIs centered at MNI:-6,8,-4 and MNI:6,10,-8 respectively. As we did not have separate hypotheses regarding the function of left and right NAcc in this context, the union of these regions was employed as a single ROI for all analyses. Regulatory regions were selected from a recent meta-analysis of the literature on emotion regulation by Kohn and colleagues (Kohn, Eickoff, Scheller, et al., 2014). Six millimeter spherical RVLPFC and LVLPFC ROIs were defined based upon the peak activation coordinates associated with cognitive emotion regulation in the left and right IFG (MNI:-42,22,-6 and MNI:50,30-8; see Table 2 in Kohn Eickoff, Scheller et al., 2014 for details). All ROIs were constructed using the Automated Anatomical Labeling (AAL) toolbox (Tzourio-Mazoyer et al., 2002) of the Wakeforest University Pickatlas (Maldjian, Laurienti, Kraft, and Burdette, 2003).

fMRI analytic paradigm:

A general linear model was defined for each participant, in which trials were modeled with separate functions corresponding to 1) the initial presentation of the trial and 2) a fixed epoch corresponding to the final 2.5 seconds preceding (and including) the participants' final response. The initial portion of the trial differs significantly between conditions, with the Confirmation and Disconfirmation conditions, but not the No Information condition, including on-screen information regarding the attitudes of another UCLA undergraduate. As parameter estimates from this portion of the trial are not directly comparable across conditions, the initial portion of each trial was therefore modeled as a parameter of no interest in the GLM. Planned comparisons were conducted on parameter estimates corresponding to the final period of each trial (i.e. the last 2.5 seconds before participant response), which we believe better corresponds to the period of participants' decision-making and response selection. Both stimulus presentation and response selection were convolved with the canonical (double-gamma) hemodynamic

response function. Four regressors of interest were modeled to the response period of the Confirmation, Disconfirmation, No Information, and Control conditions. The model also controlled for 18 motion parameters (3 translations and rotations, as well as their squares and first-order derivatives), and a junk regressor for acquisitions on which either translation exceeded 2mm or rotation exceeded 2 degrees in any direction. The time series was high-pass filtered using a cutoff period of 128s and serial autocorrelations were modeled as an AR(1) process.

Consensus bias was computed on a trial-by-trial basis as the error of estimation of a participant's consensus estimate regarding the attitude item (relative to the true mean of our larger, 197 person sample) *in the direction of the participant's own attitude on the attitude item* (acquired several days before the scan). That is, consensus bias was operationalized as | consensus estimate – true sample mean | (x-1 if consensus estimate underestimates support of own attitude). Bias values were also capped by the participant's own attitude; that is participants could not have a bias score greater than the difference between their own attitude and the sample mean. The consensus bias metric used is thus positive when participants overestimate support for their own attitudinal positions in the UCLA undergraduate population, negative when they underestimate support for their own attitudinal positions in the undergraduate population, and 0 if their estimate is accurate. Because this bias metric is sensitive to participants' actual over-estimation of support for their own attitudes, we believe it is an effective operationalization of consensus bias for the purposes of imaging research. It is conceptually similar to the 'truly false consensus effect' developed by Krueger and Clement (Krueger & Clement, 1994).

Parameter estimates were extracted from all ROIs using MarsBaR (Brett, Anton, Valabregue, and Poline, 2002) and entered into multiple regression models (see *Results* below) with participants' mean consensus bias scores (overall or condition-specific, depending upon the

model under evaluation) as the dependent variable. Additional regions whose activity correlated with between-subjects variation in consensus bias were identified by whole-brain analyses, interrogating only gray-matter voxels. Monte Carlo simulations implemented in 3dClustSim (from AFNI; Cox et al., 1996) were used to determine appropriate cluster-size thresholds (70 contiguous voxels) to ensure overall false discovery rate (FDR) of less than 0.05, when combined with a voxel-wise significance threshold of $p<0.005$ within gray-matter voxels. All results reported exceed these joint voxel-wise and cluster-extent thresholds, except as noted.

RESULTS

Behavioral effects of social information on consensus bias:

Consistent with the extensive behavioral literature on the false consensus effect, consensus bias scores were significantly greater than zero both overall and for each information condition individually ($M_{all}=12.174$, $t(27)=15.265$, $p<0.001$; $M_{Con}=19.071$, $t(27)=18.604$, $p<0.001$; $M_{NoI}=10.320$, $t(27)=9.950$, $p<0.001$; $M_{Dis}=8.272$, $t(27)=10.445$, $p<0.001$). Mean consensus bias scores were not related either to mean estimate confidence ratings ($r=0.17$, *ns*) or to mean attitude importance scores ($r=-0.185$, *ns*). Overall, there was a marginally significant inverse correlation between mean consensus bias and mean reaction time, averaging across all conditions ($r=-0.344$, $p=0.073$). Mean bias in the Confirmation condition was inversely correlated with mean reaction time to Confirmation trials ($r=-0.399$, $p=0.035$), but this relationship did not hold for the Disconfirmation or No Information conditions.

Repeated-measures analysis of variance revealed a substantial effect of information condition (Confirmation, Disconfirmation, or No Information) on participants' exhibited bias ($F(2,54)=80.580$, $p<0.001$). Participants showed greater bias in the Confirmation condition than the No Information condition ($M_{Con}=19.071$ versus $M_{NoI}=10.320$, $t(27)=9.095$, $p<0.001$).

Participants also showed significantly less bias in the Disconfirmation condition than in either the No Information condition (M_{Dis}=8.272 versus M_{NoI}=10.315, t(27)=-2.279, p=0.031) or the Confirmation condition (M_{Dis}=8.272 versus M_{Con}=19.071, t(27)=-11.509, p<0.001).

The presentation of sample information also affected participants' reaction times (F(2,54)=5.137, p=0.007). Predictably, both the Confirmation and Disconfirmation conditions resulted in longer reaction times than the No Information condition (M_{Con}=4.541 versus M_{NoI}=4.323, t(27)=3.077, p=0.005; M_{Dis}=4.522 versus M_{NoI}=4.323, t(27)=2.367, p=0.025). However, the Confirmation and Disconfirmation conditions did not differ in reaction time (M_{Con}=4.541 versus M_{Dis}=4.522, t(27)=0.274, p=0.786).

Participants' confidence in their consensus estimates was also affected by the presentation of sample information onscreen (F(2,54)=4.673, p=0.011). The Confirmation condition increased participants' confidence in their consensus estimates relative to the No Information (M_{Con}=68.761 versus M_{NoI}=66.064, t=2.662, p=0.013) and Disconfirmation (M_{Con}=68.671 versus M_{Dis}=65.676, t=2.568, p=0.016) conditions, but the Disconfirmation condition did not decrease participants' confidence in their estimates relative to No Information (M_{Dis}=65.676 versus M_{NoI}=66.064, t(27)=-0.361, p=0.720). Perceived attitude importance was not influenced by information condition (F(2,54)=1.068, p=0.351).

On average, participant response in the color judgment Control trials was not biased in favor of either color (Red or Blue) along the continuum presented (mean signed error M_{Err}=-0.048, t=-0.057,p=0.955). Participants were not terribly inaccurate in their color judgments (mean absolute error, M_{AbsErr}=9.818 out of a 100-point scale), but this error was significantly different from zero (t=24.594,p<0.001). Reaction times were shorter for the Control trials than

for the Consensus Estimation trials ($M_{ConsensusRT}$=4.460 versus $M_{ColorRT}$=3.206, paired-sample t=-8.388, p<0.001), suggesting that the color judgment task was slightly easier to perform.

Taken together, these results suggest that participants integrated the affirming and challenging information into their consensus estimates in the manner intended. The Confirmation and Disconfirmation trials took slightly longer to complete, on average, and the information provided had the expected impact on participants' demonstrated bias – enhancing and diminishing the consensus bias on Confirmation and Disconfirmation trials, respectively. Participants were slightly more confident when the sample information confirmed the normativity of their beliefs than when no information was provided. It is worth noting that, consistent with the observed consensus bias, participants were relatively confident in their estimates in all conditions. Lastly, since attitude items were assigned to experimental conditions randomly, it is reasonable that there should not be significant differences in perceived attitude importance.

Neural correlates of between-subjects variation in consensus bias

Given our assumptions about the reward and regulatory processes underlying the false consensus effect, we sought to assess whether between-subjects variation in critical regions-of-interest would predict variation in participants' observed levels of consensus bias. Specifically, as outlined above, we anticipated that reward activity in the ventromedial prefrontal cortex (VMPFC) and nucleus accumbens (NAcc) would be associated with greater consensus bias, while regulatory activity in the right ventrolateral prefrontal cortex (RVLPFC) and left ventrolateral prefrontal cortex (LVLPFC) would be associated with diminished bias.

Parameter estimates were extracted from these regions during the response period, for each information condition (Confirmation, Disconfirmation, and No Information) versus control,

and entered into a multiple regression model as predictors of between-subjects variation in consensus bias. As noted above, reaction time differed as a function of condition and was marginally inversely associated with consensus bias. In order to rule out any possible effects due simply to variation in reaction time, this variable was also included as a regressor of no interest. We first assessed whether mean task-related activity (averaging over conditions relative to control) in the regions-of-interest would significantly predict mean consensus bias (again averaging bias scores across conditions). In this model, activity in VMPFC, bilateral NAcc, RVLPFC, and LVLPFC together significantly predicted about half of the variance in participants' mean consensus bias (model $F(5,22)=4.455$; $p=0.006$; $R^2=0.503$). In addition, the neural predictors independently accounted for a significant proportion of variance in consensus bias scores: bias was positively associated with activity in NAcc ($t=2.303$, $p=0.031$, partial correlation $r=0.441$) and VMPFC ($t=2.164$, $p=0.042$, partial correlation $r=0.419$), but negatively associated with activity in RVLPFC ($t=-2.192$, $p=0.039$, partial correlation $r=-0.423$). Activity in the LVLPFC was not significantly associated with consensus bias ($t=0.287$, $p=0.777$, partial correlation $r=0.061$). These results are consistent with our predictions regarding the role of reward and regulatory processes in consensus estimation, insofar as reward-related regions (NAcc and VMPFC) were more active in participants who exhibited greater mean levels of bias, while the RVLPFC was recruited more by participants whose estimates were less biased (See Figure 2).

Similar results were uncovered when trials were analyzed in a condition-specific manner: that is, when parameter estimates extracted from the *a priori* ROIs during a given condition were used as predictors of observed bias during that condition. For No Information trials, the overall model (including NAcc, VMPFC, RVLPFC, LVLPFC, and reaction time as predictors) remained

significant (model $F(5,22)=5.636$, $p=0.002$, $R^2=0.562$). Consensus bias scores during the No Information condition were positively associated with activity in NAcc ($t=2.734$, $p=0.012$, partial correlation $r=0.504$) and VMPFC ($t=2.122$, $p=0.045$, partial correlation $r=0.412$) during this condition and negatively associated with activity in RVLPFC during this condition ($t=-3.204$, $p=0.004$, partial correlation $r=-0.564$). Again, the association between activity in the LVLPFC and consensus bias was not significant for the No Information condition ($t=0.905$, $p=0.375$, partial correlation $r=0.189$).

A comparable model also significantly predicted between-subjects variation in bias observed during the Disconfirmation condition (model $F(5,22)=4.060$, $p=0.012$, $R^2=0.414$). Consensus bias scores during the Disconfirmation condition were positively associated with activity in the VMPFC ($t=2.684$, $p=0.013$, partial correlation $r=0.488$) and marginally associated with activity in the NAcc ($t=2.019$, $p=0.055$, partial correlation $r=0.388$) but negatively associated with activity in the RVLPFC ($t=-2.572$, $p=0.017$, partial correlation $r=-0.473$). Parameter estimates from LVLPFC did not significantly predict bias in the Disconfirmation condition ($t=0.951$, $p=0.351$, partial correlation $r=0.195$). In the Confirmation condition, a multiple regression model including reaction time and ROI parameter estimates from the regions-of-interest did not significantly predict mean bias (model $F(5,22)=1.399$, $p=0.263$).

Taken together, these results demonstrate that consensus bias is positively associated with activity in regions (bilateral NAcc and VMPFC) implicated in the subjective experience of reward, and negatively associated with activity in a key regulatory region (RVLPFC) during both the Disconfirmation and No Information conditions. These are precisely the trials in which participants ought to be uncertain about the status of their own attitudes vis-à-vis those of their peers, and in which the interplay of motivated reasoning and regulation is expected to shape

observed bias. In the Confirmation condition, the same regions do not seem to predict consensus bias, perhaps because participants may take the confirmatory feedback at face value most of the time.

Whole-brain analyses were conducted to determine whether brain regions other than the *a priori* ROIs would show significant associations with between-subjects variation in consensus bias. Interestingly, this analysis revealed a cluster in the left precuneus that was positively associated with observed bias during the No Information condition (peak MNI:-3,-58,16; t=4.155; k=117). Given the role of the precuneus in retrieval processes (Kim, 2013), this result provides tentative evidence that biased retrieval may support errors of consensus estimation, even when social feedback is unavailable for direct assessment.

DISCUSSION

In this investigation, we conducted a functional neuroimaging test of a prominent social psychological account of the false consensus effect (FCE), which views consensus bias as a consequence of motivated reasoning/projection (see Marks & Miller, 1987). The results provide support for the theoretical importance of motivated projection in shaping the expression of the FCE, but also highlight participants' (limited) capacity for regulatory restraint – a factor not fully considered in previous accounts of the FCE. In the No Information and Disconfirmation conditions, established reward regions (NAcc and VMPFC) were associated with a tendency towards *greater* bias, while activity in the RVLPFC (implicated in emotion regulation and self-restraint) was *inversely* related to the consensus bias. Indeed, overall, the activity in these regions of interest accounted for almost 50% of the total between-subjects variation in consensus bias. These findings suggest, as some social psychologists have theorized (see, e.g. Crano, 1983; Sherman, Presson, & Chassin, 1984; Morrison & Matthes, 2011), that our tendency to project

our own attitudes onto others is not simply the result of the greater accessibility intrinsic to our own perspective. Indeed, these neuroimaging results are congruent with the notion that projection is (at least in part) motivated, perhaps reflecting the need to affirm the normativity of our attitudes within the broader community.

The results of the present study are also consistent with a number of cognitive and social cognitive findings concerning related phenomenon. A very similar pattern of motivated projection and regulatory restraint has been observed previously with the "belief" bias in syllogistic reasoning. The "belief" bias results when individuals are presented with a valid logical argument that results in an untrue conclusion. Consider the argument:

No addictive things are inexpensive

Some cigarettes are inexpensive

Therefore, some cigarettes are not addictive

This argument's conclusion is generally thought to be untrue, however, it is also a valid conclusion because it follows logically from the premises. Fewer than half of individuals identify this argument as logically valid (Evans, Barston, & Pollard, 1983), while showing almost perfect accuracy on trials where the participants' beliefs were not at odds with the argument's conclusion.

An fMRI study examined the "belief" bias (Goel & Dolan, 2003), including the critical trials during which participant beliefs were likely to be at odds with the validity judgment. When participants fell prey to the "belief" bias and projected their beliefs onto the validity decision, rather than preventing their own beliefs from interfering, the only region of the brain that was relatively more active was VMPFC. This is analogous to the greater VMPFC activity we observed to the extent that our participants erroneously projected their own attitudes onto the

consensus estimates of others' attitudes. In contrast, when participants overcame the "belief" bias and correctly identified the valid, but untrue, conclusions as valid, the only brain region that was relatively more active was RVLPFC. This again is analogous to our finding that reduced consensus bias was associated with RVLPFC activity.

Within social cognition, VMPFC has previously been associated with motivated social cognition (Beer & Hughes, 2010, Hughes & Beer, 2012). A number of studies also suggest that RVLPFC plays a key role in detaching from one's own perspective or existing beliefs in order to consider additional information or perspectives. For instance, when first impressions, which are notoriously difficult to change, are successfully updated, this change is associated with RVLPFC activity (Bhanji & Beer, 2013; Mende-Siedlecki, Cai, & Todorov, 2012). Additionally, in our own work, we have also observed that when adolescents change their own attitudes to be more like those of a parent or peer, there is greater activity in RVLPFC, relative to trials when less of an attitudinal shift occurred (Welborn et al., 2016).

Perhaps most compelling is a case study of a patient with damage localized to RVLPFC (Samson et al., 2005). As long as the patient had no antecedent beliefs or preferences relevant to a perspective-taking task, the patient showed perfectly preserved performance. However, when the patient had his own perspective or preference, he could not help but project this onto others, showing childlike egocentrism. If a game were being played between two teams that he did not care about personally, he could accurately assess how fans of each team would react if one of the teams scored. In contrast, if the game included the patient's own favorite team, he assumed other fans would have the same reaction as him, even if told someone was rooting for the other team.

All of the aforementioned phenomena (FCE, "belief" bias, person perception updating, and recognizing another's perspective when discrepant with our own) may be examples of a

broader phenomenon known as *naïve realism* (Ross & Ward, 1996). Naïve realism refers to the (implicit) belief that we see the world objectively and that other reasonable people should thus see it the same way we do. If they fail to see it our way, we rarely consider how our perception or understanding might be wrong or just one of several possible points-of-view. Although we often fail to overcome our own initial way of seeing things and assume others see things the same way we do, as evidenced by self-projection in the FCE, sometimes we are able to detach ourselves from our own perspective. Across these various studies, including in the current FCE findings, RVLPFC appears to play a role in overcoming naïve realism and appreciating information beyond our initial intuitive perspective.

Given that naïve realism is generally believed to be both entrenched and socially problematic, identifying neural dynamics that support even temporary detachment from this state of self-certainty and self-projection is very important. Due to naïve realism, we tend to over-estimate *others'* susceptibility to biases while underestimating our own (Pronin, Lin, & Ross, 2002; Pronin, Gilovich, & Ross, 2004). This pronounced asymmetry in perceptions of bias between self and others has been shown in a variety of important domains, including interpersonal perception (Pronin, Krueger, Savitsky, & Ross, 2001) and intergroup conflict (Robinson, Keltner, Ward, & Ross, 1995). Thus, if RVLPFC plays a central role in those occasions when naïve realism is overcome, then this may serve as a point of focus for future investigations and interventions. For instance, a recent study observed that self-control training enhanced RVLPFC responses in a region very close to the one identified in the current study (Berkman, Kahn, & Merchant, 2014). It is possible that training regimens that focus on enhanced motor self-control would also produce benefits for overcoming non-motor impulses as well, like those that must be restrained when we are under the sway of naïve realism (Berkman,

Burklund, & Lieberman, 2009).

While the results of this experiment are consistent with motivated projection as a cause of consensus bias, the diversity of function associated with the brain regions in question (especially the VMPFC), mean that other contributing factors should also be considered in future work. The VMPFC has often been implicated in self-related cognition (Jenkins & Mitchell, 2011; Tamir & Mitchell, 2010), and both the VMPFC and the nucleus accumbens have been associated with social influence processes (Welborn et al., 2016; Zaki, Schirmer, & Mitchell, 2011). Such processes are not inconsistent with motivated projection, but their involvement could be clarified by direct comparisons of self-related cognition, influence and consensus estimation in a single sample.

We should also note a number of crucial limitations regarding causal inferences based upon correlational evidence, such as the fMRI results presented in this paper. Statistical models of hemodynamic response at best reveal associations between neural activity and bias, but do not uniquely specify the causal relationships between brain regions and behaviors. In addition, there is considerable uncertainty about the timing of the psychological processes associated with consensus estimation in this paradigm. Consensus estimation trials evolved in a relatively unconstrained manner, with no clear demarcation enforced by the experimental design between the time-period during which participants were making judgments and the period of scale manipulation. Indeed, for many participants these periods may have been overlapping. Thus, it is possible that other processes, besides motivated projection and regulatory restraint, are responsible for the association between the regions specified and consensus bias. In light of previous work on the FCE and the neuroscience literature on reward and regulatory processes, we feel that an account of consensus bias in terms of motivated projection and regulatory

restraint is most consistent with the observed results. Nevertheless, other causal relationships are plausible and ought to be explicitly examined in future work. For example, activity in putative reward regions may be elicited as a response to or an effect of attitudinal projection, rather than as an antecedent cause of bias. Future research might assist in clarifying with greater precision the causal mechanisms involved in consensus bias.

The present research has explored the neural correlates of the FCE with respect to contemporary social, political and personal issues. The results of this work are consistent with social psychological accounts of consensus bias in terms of motivated reasoning, and suggest that regulatory mechanisms may offer hope for attenuating bias in the face of social feedback. Further research may profitably understand the circumstances and limits of individuals' capacities to overcome bias, as well as investigate their neural mechanisms.

REFERENCES

Bartra, O., McGuire, J.T., & Kable, J.W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage, 76,* 412-427.

Beer, J.S., & Hughs, B.L. (2010). Neural systems of social comparison and the "above-average" effect. *NeuroImage, 49,* 2671-2679.

Berkman, E.T., Burklund, L., & Lieberman, M.D. (2009). Inhibitory spillover: intentional motor inhibition produces incidental limbic inhibition via right inferior frontal cortex. *NeuroImage, 47,* 705-712.

Berkman, E.T., Kahn, L.E., & Merchant, J.S. (2014). Training-induced changes in inhibitory control network activity. *Journal of Neuroscience, 34,* 149-157.

Bhanji, J.P., & Beer, J.S. (2013). Dissociable neural modulation underlying lasting first impressions, changing your mind for the better, and changing it for the worse. *Journal of Neuroscience, 32,* 9337-9344.

Brett, M., Anton, J-L., Valabregue, R., and Poline, J-B. (2002). Region of interest analysis using an SPM toolbox [abstract] Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2-6, 2002, Sendai, Japan. Available on CD-ROM in Neuroimage 16(2).

Cohen, J.R., Bekrman, E.T., & Lieberman, M.D. (2013). Intentional and incidental self-control in ventrolateral PFC. In D.T. Stuss & R.T. Knight (Eds.) *Principles of Frontal Lobe Function* (2nd ed) (pp.417-440), New York: Oxford University Press.

Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162-173.

Crano, W. (1983). Assumed consensus of attitudes: the effect of vested interest. *Personality and Social Psychology Bulletin, 9,* 597-608.

Evans, J.B.T., Barston, J.L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11,* 295-306.

Falk, E.B., Morelli, S.A., Welborn, B.L., Dambacher, K., & Lieberman, M.D. (2013). Creating buzz: the neural correlates of message propagation. *Psychological Science, 24,* 1234-1242.

Goel, V., & Dolan, R.J. (2003). Explaining modulation of reasoning by belief. *Cognition, 87,* B11-B22.

Hartwright, C.E., Apperly, I.A., & Hansen, P.C. (2015). The special case of self-perspective inhibition in mental, but not non-mental, representation. *Neuropsychologia, 67,* 183-92.

Hughes, B.L., & Beer, J.S. (2012). Medial orbitofrontal cortex is associated with shifting decision thresholds in self-serving cognition. *NeuroImage, 61,* 889-898.

Izuma, K., Saito, D.N., Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron, 58,* 284-294.

Jenkins, A.C., & Mitchell, J.P. (2011). Medial prefrontal cortex subserves diverse forms of self-reflection. *Social Neuroscience, 6,* 211-218.

Kim, H. (2013). Differential neural activity in the recognition of old versus new events: an activation likelihood meta-analysis. *Human Brain Mapping*, *34*, 814-36.

Kohn, N., Eickhoff, S.B., Scheller, M., Laird, A.R., Fox, P.T., & Habel, U. (2014). Neural network of cognitive emotion regulation – An ALE meta-analysis and MACM analysis. *NeuroImage, 87,* 345-355.

Krueger, J., & Clement, R.W. (1994). The truly false consensus effect: an ineradicable egocentric bias in social perception. *Journal of Personality and Social Psychology, 67,* 596-610.

Maldjian, J.A., Laurienti, P.J., Burdette, J.B., and Kraft, R.A. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets.

NeuroImage, 19, 1233-1239.

Marks, G., & Miller, N. (1987). Ten years of research on the false consensus effect: an empirical and theoretical review. *Psychological Bulletin, 102,* 72-90.

Mende-Siedlecki, P., Cai. Y., & Todorov, A. (2013). The neural dynamics of updating impressions. *Social Cognitive and Affective Neuroscience, 8,* 623-631.

Morrison, K.R., & Matthes, J. (2011). Socially motivated projection: need to belong increases perceived opinion consensus on important issues. *European Journal of Social Psychology, 41,* 707-719.

Pronin, E., Gilovich, T., & Ross, L. (2004). Divergent perceptions of bias in self versus others. *Psychological Review, 111,* 781-799.

Pronin, E., Krueger, J., Savitsky, K., & Ross, L. (2001). You don't know me, but I know you: the illusion of asymmetric insight. *Journal of Personality and Social Psychology, 81,* 639-656.

Pronin, E., Lin, D.Y., & Ross, L. (2002). The bias blind spot: perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28,* 369-381.

Robinson, R.J., Keltner, D., Ward, A., & Ross, L. (1995). Actual versus assumed differences in construal: "Naïve realism" in intergroup perception and conflict. *Journal of Personality and Social Psychology, 68,* 404-417.

Ross, L., Greene, D., & House, P. (1977). The "False Consensus Effect": An Egocentric Bias in Social Perception and Attribution Processes. *Journal of Experimental Social Psychology, 13,* 279-301.

Ross, L., & Ward, A. (1996). Naïve realism: implications for social conflict and misunderstanding. In T. Brown, E. Reed, & E. Turiel (Eds.), *Values and Knowledge* (pp.103-135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Samson, D., Apperly, I.A., Kathirgamanathan, U., & Humphreys, G.W. (2005). Seeing it my way: a case of selective deficit in inhibiting self-perspective. *Brain, 128,* 1102-1111.

Sherman, S.J., Presson, C.C., & Chassin, L. (1984). Mechanisms underlying the false consensus effect: the special role of threats to the self. *Personality and Social Psychology Bulletin, 10,* 127-138.

Simon, D., Becker, M.P., Mothes-Lasch, M., Miltner, W.H., & Straube, T. (2014) Effects of social context on feedback-related activity in the human ventral striatum. *NeuroImage, 99,* 1-6.

Spunt, R.P., Falk, E.B., & Lieberman, M.D. (2010). Dissociable neural systems support retrieval

of how and why action knowledge. *Psychological Science, 21,* 1593-1598.

Spunt, R.P., & Lieberman, M.D. (2013). The busy social brain: Evidence for automaticity and control in the neural systems supporting social cognition and action understanding. *Psychological Science, 24,* 80-86.

Tamir, D.I., & Mitchell, J.P. (2010). Neural correlates of anchoring-and-adjustment during mentalizing. *Proceedings of the National Academy of the Sciences of the United States of America, 107,* 10827-10832.

Tamir, D.I., Mitchell, J.P. (2012). Disclosing information about the self is intrinsically rewarding. *Proceedings of the National Academy of Sciences, 109,* 8038-8043.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185,* 1124-1131.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage 15,* 273-89.

Welborn, B.L., Lieberman, M.D., Goldenberg, D., Fuligni, A.J., Galvan, A., & Telzer, E.H. (2016). Neural mechanisms of social influence in adolescence. *Social, Cognitive, and Affective Neuroscience, 11,* 100-9.

Zaki, J., Schirmer, J., & Mitchell, J.P. (2011). Social Influence Modulates the Neural Computation of Value. *Psychological Science, 22,* 894-900.

Figures:

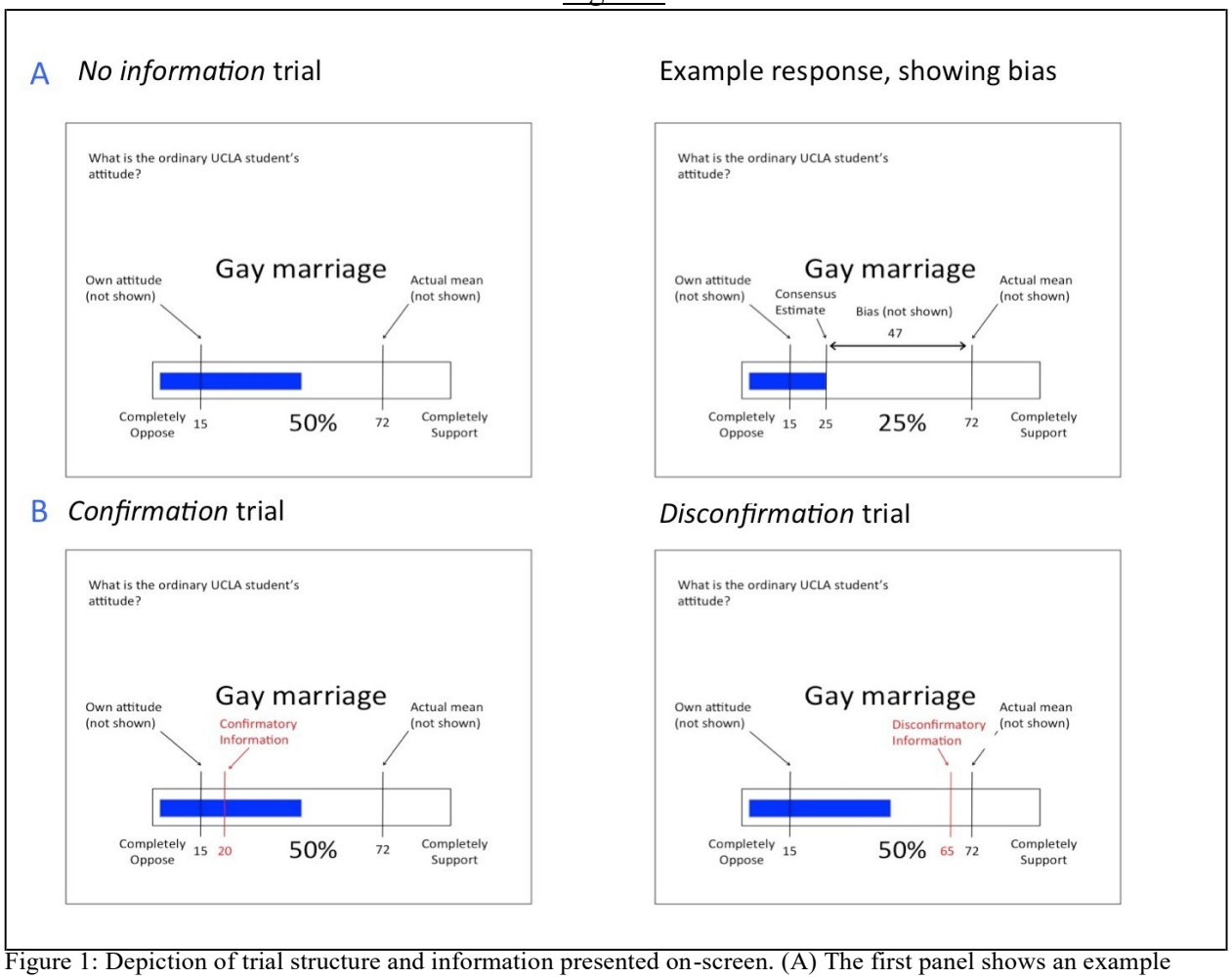

Figure 1: Depiction of trial structure and information presented on-screen. (A) The first panel shows an example screen for a No Information trial, in which a social or political attitude is presented to the participant for consensus estimation in the absence of any information ostensibly from the sample of UCLA undergraduates. In the second panel, a hypothetical response is depicted, in which a participant who opposes gay marriage selects a response that underestimates support for marriage equality in the undergraduate population. (B) Example trials from the Confirmation and Disconfirmation conditions. In the Confirmation condition, participants were presented with sample information suggesting that another undergraduate had an attitude similar to their own (no more than 5 points from their own attitude). In the Disconfirmation condition, participants were presented with sample information suggesting that another undergraduate had an attitude dissimilar to their own (at least 15 points different) and similar to the actual sample mean (within 5 points in either direction). These conditions were constrained by the experimental design to be exclusive, i.e., such that disconfirmatory information was always further from one's own attitude than confirmatory information, and always closer to the actual mean than the confirmatory information (see *Methods*).

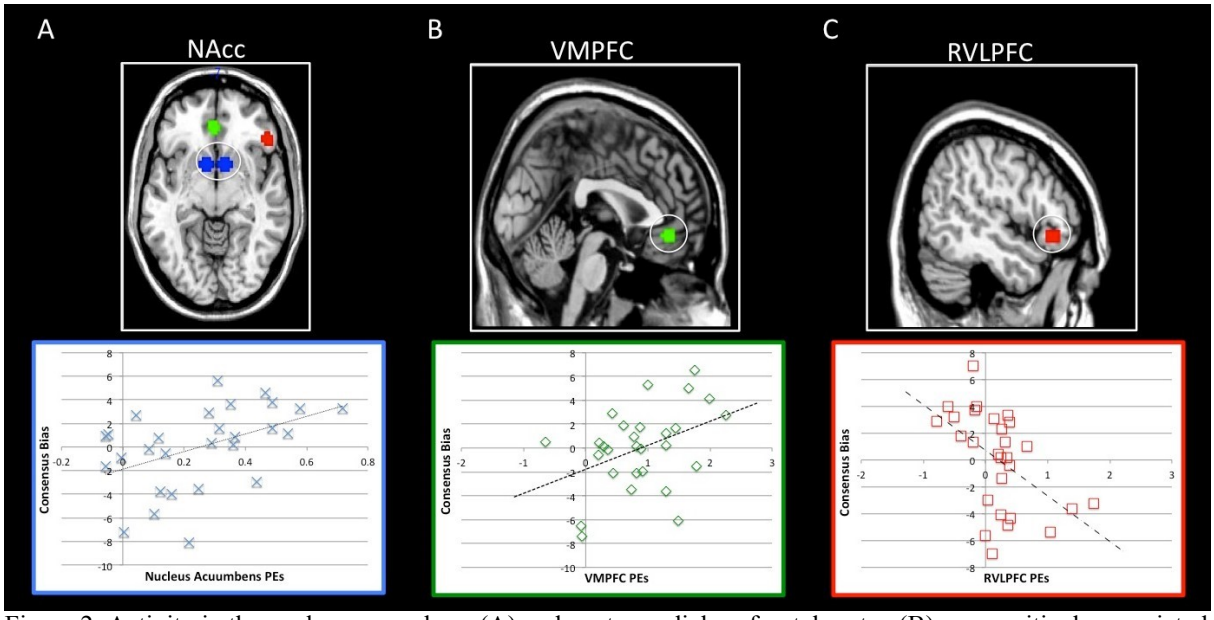

Figure 2: Activity in the nucleus accumbens (A) and ventromedial prefrontal cortex (B) was positively associated with between-subjects differences in mean consensus bias, while activity in right ventrolateral prefrontal cortex (C) was inversely associated with consensus bias. Parameter estimates are extracted from *a priori* ROIs as described above in the Methods and Results. Parameter estimates are plotted against unstandardized residual variation in consensus bias scores (that is, variation not accounted for by the other predictors).