



Published in final edited form as:

J Cogn Neurosci. 2010 September ; 22(9): 2108–2119. doi:10.1162/jocn.2009.21359.

Roles of Medial Prefrontal Cortex and Orbitofrontal Cortex in Self-evaluation

Jennifer S. Beer¹, Michael V. Lombardo², and Jamil Palacios Bhanji¹

¹University of Texas, Austin

²University of Cambridge

Abstract

Empirical investigations of the relation of frontal lobe function to self-evaluation have mostly examined the evaluation of abstract qualities in relation to self versus other people. The present research furthers our understanding of frontal lobe involvement in self-evaluation by examining two processes that have not been widely studied by neuroscientists: on-line self-evaluations and correction of systematic judgment errors that influence self-evaluation. Although people evaluate their abstract qualities, it is equally important that perform on-line evaluations to assess the success of their behavior in a particular situation. In addition, self-evaluations of task performance are sometimes overconfident because of systematic judgment errors. What role do the neural regions associated with abstract self-evaluations and decision bias play in on-line evaluation and self-evaluation bias? In this fMRI study, self-evaluation in two reasoning tasks was examined; one elicited overconfident self-evaluations of performance because of salient but misleading aspects of the task and the other was free from misleading aspects. Medial PFC (mPFC), a region associated with self-referential processing, was generally involved in on-line self-evaluations but not specific to accurate or overconfident evaluation. Orbitofrontal cortex (OFC) activity, a region associated with accurate nonsocial judgment, negatively predicted individual differences in overconfidence and was negatively associated with confidence level for incorrect trials.

INTRODUCTION

The frontal lobes have long been theorized to play an important role in self-evaluation (Stuss & Benson, 1984), but diverse empirical research has been slower to follow. Currently, neural research on the self has mostly focused on the interplay between neural systems that support self-evaluation in relation to evaluation of other people (for a review, see Uddin, Iacoboni, Lange, & Keenan, 2007; Ochsner et al., 2005). This research has shown that medial PFC (mPFC) is robustly related to semantic knowledge about the self (Uddin et al., 2007; Ochsner et al., 2005; Kelley et al., 2002). These studies provide an important foundation of knowledge; an important next step is to expand the paradigms and psychological mechanisms that are included in neural research on the self (Beer, 2007).

For example, what is the psychological mechanism through which the mPFC supports self-evaluation? One predominant explanation is that mPFC supports the representation or access to internal cues that are only available for one's own mental states that play a fundamental part in self-evaluations (Ochsner et al., 2005; Kelley et al., 2002) and may also be used in evaluating other people (Mitchell, Macrae, & Banaji, 2006; Ochsner et al., 2005). Most of the current studies have focused on evaluation of abstract information about the self (e.g., the descriptiveness of personality traits). Social psychological models of self-evaluation, particularly those focused on self-regulation, emphasize that another important self-evaluative process is evaluating one's performance in the moment (Baumeister & Heatherton, 1996). Although people might have an abstract representation of whether they are good at problem solving, self-evaluation also occurs when people evaluate their confidence in their ability to reason through a particular problem in a specific situation. In this way, on-line self-evaluation involves evaluating the self's actions, behaviors, and abilities in the moment rather than abstract representations of the self's qualities. In both cases, the self is being evaluated, but the evaluation may be focused on thinking about the self in general versus an "on-line" evaluation of the self in the moment. Although the two types of self-evaluation can be distinguished, it is likely that they may share commonalities and interact. As mentioned above, neural studies of self-evaluation of abstract qualities suggest that these evaluations involve weighting the strength of internal associations. Similarly, people may monitor internal cues to assess their on-line performance. In this way, the two processes may be computed in a similar manner. In addition, if someone has to evaluate themselves in the moment but the environment does not provide feedback, they might reference their abstract self-representations ("Am I generally good at this kind of task?"). A large body of research on the self-reference effect has established that mPFC is associated with self-evaluations of abstract qualities (Ochsner et al., 2005; Kelley et al., 2002). Does the mPFC also support on-line self-evaluation such as evaluating the self's performance on a specific task?

A second line of inquiry is examining how neural regions associated with self-evaluation are (or are not) involved in the biases that are known to characterize self-evaluation. Understanding the neural systems involved in self-evaluation biases and their correction is important because accurate on-line self-evaluation is helpful for successful self-regulation (Beer, 2007; Baumeister & Heatherton, 1996). For self-regulation purposes, individuals compare their estimations of their on-line behavior to goals and expectations. Discrepancies may motivate an adjustment of behavior or expectations of the self. However, inaccurate self-evaluation is commonly observed in healthy populations (Klayman, 1995; Taylor & Brown, 1988; Tversky & Kahneman, 1974). In the extreme, gross discrepancies between one's self-perception and one's actual behavior is a hallmark of a number of disorders (e.g., Steele, Currie, Lawrie, & Reid, 2006; Volkow et al., 1991) that have important implications for understanding treatment seeking and compliance (e.g., Aleman, Agrawal, Morgan, & David, 2006; Sanz, Constable, Lopez-Ibor, Kemp, & David, 1998). Still, very little is understood about how neural recruitment in healthy populations and neural impairments in disordered populations might relate to self-evaluation biases and their correction.

Inaccuracies in self-evaluation are known to arise for a number of reasons. For example, behavioral research has shown that people are unrealistically positive about the social

desirability of their general personal characteristics; they claim high rates of positive personal characteristics and low rates of negative personal characteristics to maintain self-worth (Taylor & Brown, 1988). This type of self-evaluation bias has been examined through the comparison of self-judgments of positive characteristics to negative characteristics and is associated with ventral ACC activity (e.g., Sharot, Riccardi, Raio, & Phelps, 2007; Moran, Macrae, Heatherton, Wyland, & Kelley, 2006).

Although overconfident assessments may sometimes occur as a self-esteem defense (e.g., Taylor & Brown, 1988), they are not always driven by emotion–regulation processes. Furthermore, overconfidence is not specific to evaluations of abstract characteristics of the self. Decades of behavioral research have shown that overconfident self-evaluations in relation to on-line behavior, such as task performance, occur in conditions where people assess themselves using information that is limited or irrelevant for evaluating their performance (for a review, see Klayman, 1995; Tversky & Kahneman, 1974). For example, people are likely to perform equally well when reasoning about forced-choice options in a number of domains (e.g., about 64% correct for reasoning about cities with higher average temperatures in July or which states have more of their population below the poverty line). However, people tend to be overconfident about their performance on some reasoning tasks (estimate 79% correct for temperature) in comparison to more accurate estimations for performance on other reasoning tasks (estimate 63% correct for poverty level) (Klayman, Soll, Gonzalez-Vallejo, & Barlas, 1999). In contrast to claims about positive versus negative personal characteristics, people do not claim to reason better in certain domains to bolster their self-esteem (i.e., reasoning better about temperature than poverty does not boost self-esteem). Instead, people reason using different kinds of information to answer questions in each domain, and these different approaches lead to different confidence estimates. In both cases, participants do not know the exact average temperature in July for most cities or the exact percentage of each state’s population below the poverty. Therefore, this task does not measure evaluations of confidence in one’s ability to retrieve or remember information they have directly learned. Instead, participants have to draw on whatever information they deem helpful for reasoning through the forced-choice options. Information that appears relevant for reasoning about the temperature questions is perceived as more readily available (e.g., geographical location of the cities, whether the city attracts tourists, etc.) than for the poverty questions. As in many other domains of judgment, available information often gets overemphasized when judging one’s performance on a task and leads to overconfidence (Tversky & Kahneman, 1974). In other words, self-evaluations of performance in the domains of temperature and state poverty levels are proxies for two kinds of self-evaluations—self-evaluations in which participants erroneously believe their reasoning performance is bolstered by an increased presence of retrievable facts and self-evaluations of reasoning ability in a context where facts may not seem as salient. Participants tend to systematically boost their confidence estimates because they believe the presence of the easily available information strengthens their performance compared with reasoning in a domain that does not lend itself to easily available sources of information. However, as mentioned above, participants make a systematic judgment error by overemphasizing the importance of their retrieved information because performance does not significantly differ across the reasoning tasks (Klayman et al., 1999).

What neural regions might be expected to mediate biased self-evaluation that may arise from systematic judgment errors? Very little is known about the neural mechanisms of self-perceptual biases or accuracy (Beer, 2007). The relation between mPFC and abstract self-evaluation suggests that this region may be important for mediating over-confidence. For example, when people are asked to evaluate their on-line behavior but do not feel the situation provides enough information, they may draw on how they generally view themselves to estimate their on-line behavior. For example, if a person is trying to ascertain how they are doing on a task but receiving no feedback, he might draw on his general representation of his abilities to make the on-line evaluation. Overconfidence may be avoided when abstract self-representations are used for the on-line evaluation because the very process of having to look outside the situation for information about the self should lower confidence. Other research indicates that the OFC might be involved in avoiding overconfident bias. Patients with selective OFC damage are overconfident in their assessment of their social competence in comparison to healthy control participants and patients with lateral frontal damage (Beer, John, Scabini, & Knight, 2006). Therefore, overconfident self-views may be associated with a failure or suppression of OFC recruitment.

The present research examines neural activity in relation to two understudied processes: on-line self-evaluation and self-evaluation bias. Participants reasoned about forced-choice options in two domains (temperatures and poverty levels). After reasoning about each forced-choice pair, participants rated their confidence that their reasoning resulted in a correct response. Previous research has shown that there are no significant differences in performance across domains, but one domain (temperature) is associated with overconfident self-evaluations whereas self-evaluations for reasoning success in the other domain (poverty) tend to be more accurate (Klayman et al., 1999). Therefore, this paradigm is useful for examining neural processes associated with making general on-line self-evaluations as well as biased on-line self-evaluations. The neural activity associated with making on-line self-evaluation was examined through a conjunctive analysis of significant activation across confidence estimates for both reasoning tasks. If mPFC is associated with on-line self-evaluation, then it should show significant change across confidence estimates. Overconfident self-perception was examined by comparing the condition of overconfident self-perception to the condition of relatively more accurate self-perception. If mPFC mediates self-evaluation bias, then it should be significantly related to overconfident self-beliefs when compared with accurate self-beliefs. Alternatively, overconfident self-evaluation may reflect a failure to recruit OFC.

METHODS

Participants

Sixteen right-handed participants (9 women; age, $M = 21.7$ years, $SD = 5.3$ years) were recruited in compliance with the University of California, Davis, human subjects regulations and were compensated \$10/hr for their participation. All participants were screened for medications or psychological and/or neurological conditions that might influence the measurement of CBF.

Behavioral Paradigm

Participants made self-evaluations of their reasoning ability in a reasoning task used in previous research (Klayman et al., 1999). On each trial, participants had to reason through a forced-choice problem and then rate their confidence in their reasoning. As in previous research, participants did not know the exact value of each forced-choice option but had to reason about which option was most likely (Klayman, 1995; Tversky & Kahneman, 1974). On the basis of previous research and pilot testing, two reasoning domains were selected that were similar in difficulty but differed in their elicitation of overconfident compared with accurate self-evaluations of reasoning ability: temperature (average July city temperatures) and poverty (percentage of state population below poverty level) (Klayman et al., 1999). Pilot testing showed that our population of participants did not know exact average July temperatures of various cities nor did they know exact percentages of state populations under the poverty level. Instead, participants used whatever information they could to reason about which city might have a higher average July temperature or which state might have more people at the poverty level. As expected, participants tended to assume that they were more successful at reasoning about the temperature questions because they found it easier to retrieve information they believed to be relevant for that task (i.e., geographical location, tourist attractions, etc.), whereas relevant sources of information felt less available for the poverty questions.

In each trial, participants were first presented with forced-choice options from either the temperature or the poverty condition for 4000 msec. In the temperature condition, participants were asked, “Which of these tourist cities had a warmer daily high temperature in July, on average?” and used a button box to indicate their choice from two options (e.g., Seoul, Athens). In the poverty condition, participants were asked, “Which of these states had a higher percentage of its population below the federal poverty line in 2003?” and given two U.S. states to choose from (e.g., Kansas, Montana). After making a choice, participants were presented with a fixation screen indicating that they should clear their minds. These fixation screens were jittered with lengths of 2 sec (50%), 4 sec (25%), or 6 sec (25%). The duration of the fixation point screens was jittered so that activity in relation to the question and the confidence estimate could be analyzed independently (Donaldson, Petersen, Ollinger, & Buckner, 2001). Participants were then presented with a confidence estimate screen (2000 msec) that asked “How confident are you that you chose the correct answer?” and provided response options in 5% increments from 50% (chance) to 95%. Participants responded using button boxes (each hand had a five-button box). Increments from 50% to 95% were used because participants only had access to 10 buttons and pilot testing showed that participants rarely used the 100% option but did make use of the 50% chance option. The confidence estimate screen was followed by a fixation screen that was randomly jittered in the same manner as the first fixation screens. Participants were not given feedback on whether their answer was correct. Participants completed five runs each consisting of 25 trials of each of the temperature and poverty conditions (125 trials for each condition total). The temperature and poverty trials were randomly intermixed within a run, and runs lasted about 10 min and 52 sec.

For all runs, stimuli were projected onto a screen mounted on the bed of the scanner. Participants' head motion was limited using foam padding. Stimulus presentation and response collection was controlled by the program Presentation running on a Windows 98 computer.

MRI Data Acquisition

All images were collected on a 1.5-T GE Signa scanner at the University of California, Davis, Imaging Research Center. Functional images were acquired with a gradient-echo EPI sequence (repetition time = 2000 msec, echo time = 40 msec, field of view = 220, 64×64 matrix, voxel size = $3.444 \times 3.44 \times 5$ mm) with each volume consisting of 24 oblique axial slices which were tilted -15° from the AC-PC line to preserve whole-brain coverage while optimizing coverage of the OFC. Both coplanar and high-resolution T1-weighted images were also acquired from each subject so that data could be normalized to the Montreal Neurological Institute atlas space. Structural and coplanar images were normalized to the T1 templates and the parameters from the coplanar normalization were used to normalize the functional images. The normalization algorithm used a 12-parameter affine transformation together with a nonlinear transformation involving cosine basis functions and resampled the volumes to 2-mm cubic voxels.

MRI Data Analysis

All statistical analyses were conducted using SPM2 (Wellcome Department of Cognitive Neurology). Functional images were reconstructed from k -space using a linear time interpolation algorithm to double the effective sampling rate. Image volumes were corrected for slice-timing skew using temporal sinc interpolation, corrected for movement using rigid-body transformation parameters, and then smoothed with an 8-mm FWHM Gaussian kernel. To remove drifts within sessions, a high-pass filter with a cutoff period of 128 sec was applied.

A fixed-effects analysis was used to model event-related responses for each participant. The model examined responses related to reasoning (2: poverty and temperature), confidence estimate (4: poverty confidence estimate for incorrect judgments, poverty confidence estimate for correct judgments, temperature confidence estimate for correct judgments, and temperature confidence estimate for incorrect judgments), and parametric modulation of the four confidence estimate regressors. Regressors were modeled as events with a canonical hemodynamic response function with a temporal derivative. The fixation screens in between the reasoning and the confidence estimate probes were entered as a covariate of no interest to avoid possible confounds from subjects thinking about either the reasoning question they had just completed or the upcoming confidence estimate. The fixation screens following the confidence estimates were used as an estimate of baseline. A general linear model analysis was then used to create contrast images for each participant summarizing differences of interest.

Contrasts from each participant were used in a second-level analyses treating participants as a random effects. Group average SPM $\{t\}$ maps were created to contrast (1) the poverty confidence estimate condition (collapsed across incorrect and correct) and (2) the

temperature confidence estimate condition (collapsed across incorrect and correct) with the baseline condition and were thresholded at $p < .005$ with an extent threshold of 15 voxels. These maps were used in further analysis in two ways. First, a conjunction analysis using the minimum statistic compared with the conjunction null (Nichols, Brett, Andersson, Wager, & Poline, 2005) was conducted to examine neural commonalities across confidence estimates. In particular, it was predicted that a region of the mPFC associated with self-referential processing might be associated with confidence estimates across reasoning task conditions. Previous studies of self-reference have found that differences in mPFC usually reflect differences in deactivation relative to baseline (rather than differential activation; e.g., Moran et al., 2006; Macrae, Moran, Heatherton, Banfield, & Kelley, 2004; Kelley et al., 2002). It should be noted that the region of mPFC found in these studies of self-referential personality trait judgments is distinct from the dorsal region of mPFC discussed in relation to default self-referential mode models of brain activation (e.g., Gusnard & Raichle, 2001). On the basis of the work of Kelley et al. (2002), the conjunction analysis examined common voxels of activation as well as mPFC deactivation generally associated with confidence estimates across the temperature and poverty conditions. In other words, this analysis was performed by computing the intersection of the maps of significant activity associated with the “temperature confidence estimate > baseline” contrast and the “poverty confidence estimate > baseline” contrast.

Second, the group average SPM $\{t\}$ maps that directly contrasted the temperature confidence estimate (2: incorrect and correct) and poverty confidence estimate (2: incorrect and correct) conditions only considered areas that were significantly activated above baseline or the hypothesized mPFC deactivation below baseline for both or one of the confidence estimate conditions. Results from parametric modulation of confidence estimates were restricted to neural regions that differentiated confidence estimates across conditions. As above, maps were thresholded at $p < .005$ with an extent threshold of 15 voxels. Masking and ROI parameter estimates were computed using the Marsbar tool-box (Brett, Anton, Valabregue, & Poline, 2002). Maxima are reported in ICMB152 coordinates as in SPM2. Finally, group average SPM $\{t\}$ maps were created to contrast (1) the poverty reasoning condition and (2) the temperature reasoning conditions and were thresholded at $p < .005$ with an extent threshold of 15 voxels. This analysis examined differences in neural activity associated with performing the different reasoning tasks.

RESULTS

Behavioral Performance Comparable across Domains but Overconfidence Is Domain Specific

Consistent with previous research, participants were over-confident in their assessments of their reasoning performance in the temperature condition and accurate in their assessments of their reasoning performance in the poverty condition despite performing equally across the reasoning tasks (Klayman et al., 1999). As in Klayman et al. (1999), comparable measures of reasoning performance and confidence estimates were created (a) by calculating actual performance as the percentage of answers that were correct in a given condition and (b) by averaging confidence percentage estimates within a condition. In other words,

comparisons between actual performance and confidence estimates within a condition were conducted by comparing the percentage of questions answered correctly to the average percentage of confidence level for that condition. In this way, a participant who answered about 60% of the questions correctly and, on average, reported a confidence level of 60% is considered to be relatively more accurate in their self-evaluations than a participant who answered 60% of the questions correctly and, on average, reported a confidence level of 80%.

Participants' reasoning performance in the temperature and poverty conditions did not significantly differ across the conditions (actual performance: temperature, $M = 62.1\%$, $SD = 5.6\%$; poverty, $M = 65.9\%$, $SD = 8.3\%$), $t(15) = 1.30$, ns , but did exceed chance (one-sample t test), temperature, $t(15) = 8.6$, $p < .05$; poverty, $t(15) = 7.7$, $p < .05$. The two domains did not differ in actual difficulty, and participants performed the tasks significantly better than if they were guessing.

However, participants' confidence estimates were significantly different across conditions (confidence estimate: temperature, $M = 73.3\%$, $SD = 5.3\%$; poverty, $M = 70.5\%$, $SD = 6.3\%$), $t(15) = 3.1$, $p < .05$. Furthermore, participants were overconfident about their reasoning ability in the temperature condition because their confidence estimates significantly differed from actual performance, $t(15) = 5.1$, $p < .05$, but were accurate in the poverty condition because there was no significant difference between their actual performance and confidence estimate, $t(15) = 1.9$, $p > .05$. In addition, the degree of difference between actual performance and confidence estimate significantly differed across conditions (temperature, $M = 11.4\%$, $SD = 8.6\%$; poverty, $M = 3.5\%$, $SD = 8.4\%$), $t(15) = 3.6$, $p < .05$. Confidence estimates were almost always somewhat greater than actual performance in the temperature condition. In contrast, confidence estimates in the poverty condition were centered closer to "0," that is, very little discrepancy between actual performance and confidence estimate.

Follow-up analyses clarified that (a) the average confidence did not predict actual performance in either domain, (b) the participants were more confident on trials they got correct than those they got incorrect in both domains, and (c) the discrepancy between confidence and actual performance was present for both correct and incorrect trials in the temperature domain. Overconfidence could not merely be equated with high confidence in either domain; there was no correlation between average confidence and actual performance (poverty, $r = -.17$, $p < .05$; temperature, $r = -.01$, $p < .05$). Although no feedback was given, participants demonstrated sensitivity to which trials they got correct. Confidence estimates were significantly greater for correct trials than for incorrect trials in both domains: poverty confidence correct trials, $M = 71.3\%$, $SD = 5.4\%$; poverty confidence incorrect trials, $M = 66.2\%$, $SD = 6.2\%$, $t(15) = 5.02$, $p < .05$; temperature confidence correct trials, $M = 75.1\%$, $SD = 5.4\%$; temperature confidence incorrect trials, $M = 68.9\%$, $SD = 6.7\%$, $t(15) = 7.3$, $p < .05$. Finally, the discrepancy between confidence and actual performance was significant for both incorrect, $t(15) = 18.7$, $p < .05$, and correct trials, $t(15) = 5.9$, $p < .05$, in the temperature condition.

RTs were significantly different across conditions during the reasoning task (temperature, $M = 2386$ msec, $SD = 389$ msec; poverty, $M = 2286$ msec, $SD = 412$ msec), $t(15) = 2.5, p < .05$, but were not significantly different across domains for the confidence estimates (temperature, $M = 962$ msec, $SD = 183$ msec; poverty, $M = 1009$ msec, $SD = 213$ msec), $t(15) = 1.9, p > .05$. Participants took longer to reason in the temperature condition but showed no significant difference in the amount of time they took to make confidence estimates for each task.

mPFC Deactivation Occurs for On-line Self-evaluations Regardless of Domain

Activity in relation to on-line self-evaluations, that is, confidence estimates irrespective of reasoning task, was examined through a conjunction analysis between (a) the contrast of the temperature confidence estimate condition in relation to baseline and (b) the contrast of the poverty confidence estimate condition in relation to baseline. Similar to the mPFC region found in studies of abstract self-evaluation (e.g., 10 52 2, Kelley et al., 2002; -4 58 -12, Lieberman, Jarcho, & Satpute, 2004; 0, 50, 8 and -9, 50, 0, Macrae et al., 2004; -12 50 -4, Vogeley et al., 2001; -3, 47, 0, Moran et al., 2006; -4 68 -12, Ruby & Decety, 2003), the mPFC (-6 52 -12) significantly deactivated in relation to baseline for confidence estimations across condition (see Figure 1 and Table 1), $t(15) = -3.33, p < .05$ and $t(15) = -3.6, p < .05$ for temperature and poverty, respectively. mPFC deactivation was not significantly different between the temperature and the poverty confidence estimate conditions, $t(15) = -.63, p > .05$. In addition, significant activation was found in the superior and middle frontal gyri, the SMA, the inferior parietal cortex, and the lingual gyrus.

Orbitofrontal Cortex Activity Associated with Attenuating Overconfident Bias

Previous research has shown that orbitofrontal damage is associated with overconfident self-evaluations of task performance (Beer et al., 2006; Beer, Heerey, Keltner, Scabini, & Knight, 2003). This research suggests that OFC activation should be negatively correlated with overconfident self-evaluations. ROIs within OFC that might relate to over-confidence were identified by comparing the temperature confidence estimate condition to the poverty confidence estimate condition. This contrast revealed several activations in the OFC (see Table 2). Further analyses revealed that (a) one orbitofrontal region (-6 26 -12) negatively predicted overconfidence at the individual level (i.e., predicted a discrepancy between an individual's actual performance and an average confidence estimate) and (b) one orbitofrontal region (20 30 -24) was parametrically related to lower confidence estimates for incorrect trials. OFC played a role in overconfidence by predicting individuals' degree of over-confidence in the temperature condition and by predicting calibration of confidence after incorrect trials in both conditions.

The magnitude of each participant's overconfidence bias (the behavioral difference between confidence estimate and actual performance) was entered as a regressor for the contrast between temperature confidence estimates and baseline (only significant regions from the direct contrast between temperature and poverty confidence estimates were considered). This analysis showed a significant negative correlation ($r = -.66, p < .05$) between overconfidence bias in the temperature condition and OFC activity (Brodmann's area [BA] 11, peak at 8, 28, -10, $p < .005$; see Figure 2B and C). This region was significantly activated in

comparison to baseline, $t(15) = 3.6, p < .05$. In the poverty confidence estimate condition, participants did not tend to be overconfident nor did this region of OFC activate significantly differently than baseline, $t(15) = 1.13, p > .05$ (see Figure 2A). However, for comparison purposes, a correlation was conducted using an index of overconfidence bias and OFC parameter estimates from the poverty confidence estimate condition. This correlation was not statistically significant ($r = -.39, p > .05$) and tended toward significant difference from the correlation in the temperature condition ($z = -1.46, p = .07$). In the condition designed to elicit overconfidence, participants who were most likely to recruit their OFC were the participants who were most likely to avoid overconfident self-evaluations in that condition.

Another region of OFC identified in the direct contrast between temperature and poverty confidence estimate (BA 11, peak at 20 30 -24; see Figure 3A) was significantly associated with negative increments in confidence on a trial-by-trial basis in the temperature condition (BA 11, peak = 22, 28, -22, $t = 3.54$; see Figure 3B and C). Further analysis of this region's parameter estimates across conditions revealed that this effect was driven by modulation of confidence estimates following incorrect trials in the temperature condition (see Figure 3C). The temperature confidence incorrect condition showed a stronger parametric effect compared with the temperature confidence correct parametric regressor, $t(15) = 2.2, p < .05$, and tended toward significant difference compared with the poverty confidence incorrect parametric regressor, $t(15) = 1.9, p = .07$. This region's relation to confidence level did not significantly differ across the regressors from the poverty condition, $t(15) = .70, p > .05$. Furthermore, the temperature confidence estimate for incorrect trials was the only beta that was significantly different than zero, $t(15) = 4.5, p < .05$; temperature confidence correct, $t(15) = -1.35$; poverty confidence incorrect, $t(15) = 1.6$; poverty confidence correct, $t(15) = 1.2$. This region of OFC was down modulated by increments of overconfidence on incorrect trials and tended to show its strongest parametric relation in the temperature confidence incorrect condition.

The contrast between confidence estimates in the poverty condition and temperature condition found significant activation in the frontal lobes, parietal cortex, fusiform gyrus, lingual gyrus, and visual areas (see Table 2).

Reasoning in the Overconfident Domain Engages Regions Associated with Memory Retrieval

Although the purpose of the study was to examine neural activation in relation to on-line self-evaluation rather than actual performance on the self-evaluation task, exploratory analyses contrasted reasoning in the temperature domain to the poverty domain (see Table 3). This analysis showed significant activation in regions associated with memory retrieval effort including anterior pFC (BA 8), bilateral pFC (BA 6/9/44/45/46), and left parietal cortex (BA 7/40) as well as temporal cortex (BA 20/37) (see Figure 4; Skinner & Fernandes, 2007; Wheeler & Buckner, 2004; Nyberg, Cabeza, & Tulving, 1996). Conversely, reasoning in the poverty domain was associated with temporal cortex regions (BA 21/23) as well as cingulate and paracingulate regions (BA 23, 10).

DISCUSSION

The current study moves beyond the abstract self-evaluation paradigms typically used in neural investigations of self-processing and examines the neural systems that support on-line self-evaluations and their biases. Similar to the robust relations between mPFC and self-evaluations of general personality traits (e.g., Moran et al., 2006; Ochsner et al., 2005; Macrae et al., 2004; Kelley et al., 2002), significant mPFC changes were associated with on-line self-evaluations of task performance. However, mPFC activity did not predict self-evaluation overconfidence (i.e., a discrepancy between actual performance and confidence). Instead, OFC activity was negatively associated with overconfidence. Consistent with lesion research (Beer et al., 2006), OFC activity was associated with suppressing overconfident on-line self-evaluations at the individual and trial level of analysis. These findings have a number of implications for understanding the roles of the mPFC and OFC in self-evaluation processes.

mPFC and On-line Self-evaluation

The current study found that confidence estimates across conditions modulated a region of mPFC identified in previous studies of self-evaluation (Moran et al., 2006; Lieberman et al., 2004; Macrae et al., 2004; Ruby & Decety, 2003; Kelley et al., 2002; Vogeley et al., 2001). Although future research is needed to more robustly understand this finding, it raises two possibilities for the role of the mPFC in on-line self-evaluation. An integration of findings from the current study and previous neural research suggests that the mPFC supports a psychological process that is (a) common to self-evaluation of abstract traits and on-line behavior or (b) that abstract self-representations may be factored into on-line self-evaluations under certain conditions.

The relation between medial PFC and abstract self-representations is theorized to reflect the medial PFC's role in representing or accessing relevant internal cues such as whether personality traits are strongly or weakly associated with self (e.g., Moran et al., 2006; Macrae et al., 2004; Kelley et al., 2002). This explanation is consistent with the view in the field of judgment science that on-line self-evaluations of confidence are made by monitoring the strength of internal signals generated by reasoning efforts (Klayman, 1995; Tversky & Kahneman, 1974). Just as the medial PFC is important for monitoring internal signals about the association strength between "self" and a personality trait, it may be important for monitoring the strength of internal signals associated with one's reasoning process about each forced-choice option.

A second possibility is that the medial PFC changes in the current study reflect on-line self-evaluation that partly relies on abstract self-representations. Participants did not know the exact values of the forced-choice options and did not receive feedback on whether they had answered correctly. In the absence of explicit feedback as a mechanism for estimating task performance in both conditions, the participants may have looked for additional information sources to make their confidence estimates. In this case, participants may have partially factored in general representations of their reasoning abilities. This possibility is consistent with a metamemory study that found an association between mPFC deactivation and low self-confidence in performance regardless of whether performance was correct ($-3.57 - 12$;

Chua, Schachter, Rand-Giovannetti, & Sperling, 2006). In other words, this region is associated with general uncertainty and is not modulated by whether that uncertainty is warranted by poor performance. The metamemory study used a task that is distinct from the task in the current study. Participants had to determine whether they could recollect stimuli they had recently learned (Chua et al., 2006). In contrast, participants in the current study had to estimate how well they had reasoned through choices for which they had not learned the exact information; they were not estimating their ability to remember a specific fact. The consistent relation between mPFC and low levels of confidence regardless of actual performance across a diverse set of tasks suggests that the medial prefrontal changes generalize to conditions of uncertainty in estimating on-line behavior (rather than something specific to a particular task or discrepancy from actual performance). Therefore, in light of the research on mPFC and abstract self-representation (e.g., Moran et al., 2006; Ochsner et al., 2005; Macrae et al., 2004; Kelley et al., 2002), the mPFC deactivation associated with on-line self-evaluation may reflect people's need to draw on general representations of self ("Am I generally good at this kind of task?") when they do not feel they have enough information from the task itself to judge their performance.

Orbitofrontal Cortex Activation Attenuates Overconfident On-line Self-evaluations for Individuals and Incorrect Trials

The current study found that OFC activation predicted who was likely to be less biased in the temperature domain as well as predicted appropriate confidence calibration after incorrect trials in both domains. Previous neural research has shown that OFC is associated with (a) accurate evaluations in some domains but not others (Beer et al., 2003, 2006), (b) individual differences in accuracy when accuracy requires the suppression of salient but irrelevant information (e.g., DeMartino, Kumaran, Seymour, & Dolan, 2006; Beer, Shimamura, & Knight, 2004), and (c) parametric modulation of accuracy on a trial-by-trial basis in metamemory tasks (Schnyer, Nicholls, & Verfaellie, 2005). For example, OFC is associated with accurate (i.e., rational) gambling decisions when they require the suppression of salient but irrelevant valenced aspects of the decision options. Individual differences in OFC activity predict increased rationality, that is, less susceptibility to irrelevant information about guaranteed wins or losses for gambles that are monetarily equivalent (DeMartino et al., 2006). OFC activity also parametrically tracks accurate predictions of one's ability to recall recently learned information (Schnyer et al., 2005).

In the current study, overconfidence should have been especially elicited when participants overemphasized the value of their information retrieval efforts for their successful task performance. The temperature reasoning condition should have been associated with greater memory retrieval efforts than the poverty reasoning condition. Consistent with this interpretation, participants took longer to make a decision in the temperature condition and activated neural regions that have been associated with memory retrieval in other paradigms. In contrast, the reasoning in the poverty condition occurred more quickly and elicited activation in neural regions associated with the "default mode of activation" (Gusnard & Raichle, 2001), suggesting that reasoning judgments may have been characterized by some kind of default heuristic and less by memory retrieval efforts. In this case, the OFC region that was modulated by individual differences in bias in the temperature condition may have

reflected how much individuals strove to calibrate the value of the retrieved information (“Did all of those facts really help me answer the question?”). Such calibration should have played less of a role in the poverty condition that was characterized by a different reasoning approach. However, as the behavioral results show, participants were somewhat sensitive to when they reasoned incorrectly in both conditions. The OFC region that exhibited down modulation by confidence levels after incorrect trials may therefore reflect trial-by-trial success at confidence calibration when performance is poor.

Conclusion

More research is needed to fully understand the brain systems that support the collection of psychological processes that shape the self beyond abstract representation (e.g., Beer, 2007; Cunningham, Raye, & Johnson, 2005). The present research deepens our understanding of frontal lobe involvement in on-line self-evaluation as well as self-evaluation bias that arises from systematic judgment errors. Regions of mPFC that have previously been associated with abstract self-evaluation were engaged by tasks requiring on-line self-evaluations in the current study. Future research is needed to more fully understand the role of mPFC in on-line self-evaluation. Studies that include a self-reference localizer task and an on-line self-evaluation task or contrast conditions of on-line self-evaluation that explicitly differ in certainty are needed to strengthen the claim that the region of mPFC found in this current study truly relates to both kinds of self-evaluation and predicts certainty in on-line self-evaluation. Another remaining question is whether the mPFC activates for on-line self-evaluation tasks because abstract and on-line evaluations share a common psychological mechanism or because abstract self-representations may be used for on-line evaluation in situations of uncertainty.

Furthermore, more research needs to examine the systematic biases that affect self-evaluation at the abstract and on-line level of analysis. The current study found that subregions within the OFC tracked bias across individuals and within incorrect trials, which is consistent with its role in other paradigms (e.g., DeMartino et al., 2006; Schnyer et al., 2005). Future research is needed to better understand the multiple roles that OFC plays in attenuating bias. Another line of inquiry might more systematically examine self-evaluation bias arising from self-esteem defense compared with systematic judgments errors. Although other studies have associated bias with executive function regions such as ACC (Sharot et al., 2007; Moran et al., 2006), the current study suggests that self-evaluation biases arising from systematic judgment errors may reflect a failure to engage executive function regions such as the OFC.

References

- Aleman A, Agrawal N, Morgan KD, David AS. Insight into psychosis and neuropsychological function meta-analysis. *British Journal of Psychiatry*. 2006; 189:204–212. [PubMed: 16946354]
- Baumeister RF, Heatherton TA. Self-regulation failure: An overview. *Psychological Inquiry*. 1996; 7:1–15.
- Beer JS. The default self: Feeling good or being right? *Trends in Cognitive Sciences*. 2007; 11:187–189. [PubMed: 17347027]

- Beer JS, Heerey EH, Keltner D, Scabini D, Knight RT. The regulatory function of self-conscious emotion: Insights from patients with orbitofrontal damage. *Journal of Personality and Social Psychology*. 2003; 85:594–604. [PubMed: 14561114]
- Beer JS, John OP, Scabini D, Knight RT. Orbitofrontal cortex and social behavior: Integrating self-monitoring and emotion–cognition interactions. *Journal of Cognitive Neuroscience*. 2006; 18:871–880. [PubMed: 16839295]
- Beer, JS.; Shimamura, AP.; Knight, RT. Frontal lobe contributions to executive control of cognitive and social behavior. In: Gazzaniga, MS., editor. *The newest cognitive neurosciences*. 3. Cambridge, MA: MIT Press; 2004. p. 1091-1104.
- Brett, M.; Anton, JL.; Valabregue, R.; Poline, JB. Region of interest analysis using an SPM toolbox. Presented at the 8th International Conference on Functional Mapping of the Human Brain; Sendai, Japan. 2002. Available on CD-ROM in *Neuroimage*
- Chua EF, Schacter DL, Rand-Giovannetti E, Sperling RA. Understanding metamemory: Neural correlates of the cognitive process and subjective level of confidence in recognition memory. *Neuroimage*. 2006; 29:1150–1160. [PubMed: 16303318]
- Cunningham WA, Raye CL, Johnson MK. Neural correlates of evaluation associated with promotion and prevention regulatory focus. *Cognitive, Affective & Behavioral Neuroscience*. 2005; 5:202–211.
- DeMartino B, Kumaran D, Seymour B, Dolan RJ. Frames, biases, and rational decision-making in the human brain. *Science*. 2006; 313:684–687. [PubMed: 16888142]
- Donaldson DI, Petersen SE, Ollinger JM, Buckner RL. Dissociating state and item components of recognition memory using fMRI. *Neuroimage*. 2001; 13:129–142. [PubMed: 11133316]
- Gusnard DA, Raichle ME. Searching for a baseline: Functional imaging and the resting human brain. *Nature Reviews Neuroscience*. 2001; 2:685–693.
- Kelley WM, Macrae CN, Wyland CL, Caglar S, Inati S, Heatherton TF. Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience*. 2002; 14:785–794. [PubMed: 12167262]
- Klayman, J. Varieties of confirmation bias. In: Busemeyer, J.; Hastie, R.; Medin, DL., editors. *Psychology of learning and motivation*. Vol. 32. New York: Academic Press; 1995. p. 365-418.
- Klayman J, Soll JB, Gonzalez-Vallejo C, Barlas S. Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*. 1999; 79:216–247. [PubMed: 10471362]
- Lieberman MD, Jarcho JM, Satpute AB. Evidence-based and intuition-based self-knowledge: An fMRI study. *Journal of Personality and Social Psychology*. 2004; 87:421–435. [PubMed: 15491269]
- Macrae CN, Moran JM, Heatherton TF, Banfield J, Kelley WM. Medial prefrontal activity predicts memory for self. *Cerebral Cortex*. 2004; 14:647–654. [PubMed: 15084488]
- Mitchell J, Macrae CN, Banaji M. Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*. 2006; 50:655–663. [PubMed: 16701214]
- Moran JM, Macrae CN, Heatherton TF, Wyland CL, Kelley WM. Neuroanatomical evidence for distinct cognitive and affective components of self. *Journal of Cognitive Neuroscience*. 2006; 18:1586–1594. [PubMed: 16989558]
- Nichols T, Brett M, Andersson J, Wager T, Poline JB. Valid conjunction inference with the minimum statistic. *Neuroimage*. 2005; 25:653–660. [PubMed: 15808966]
- Nyberg L, Cabeza R, Tulving E. PET studies of encoding and retrieval: The HERA model. *Psychonomic Bulletin & Review*. 1996; 3:135–148. [PubMed: 24213861]
- Ochsner KN, Beer JS, Robertson EA, Cooper J, Gabrieli JDE, Kihlstrom JF, et al. The neural correlates of direct and reflected self-knowledge. *Neuroimage*. 2005; 28:797–814. [PubMed: 16290016]
- Ruby P, Decety J. What you believe versus what you think they believe: A neuroimaging study of conceptual perspective-taking. *European Journal of Neuroscience*. 2003; 17:2475–2480. [PubMed: 12814380]
- Sanz M, Constable G, Lopez-Ibor I, Kemp R, David AS. A comparative study of insight scales and their relationship to psychopathological and clinical variables. *Psychological Medicine*. 1998; 28:437–466. [PubMed: 9572100]

- Schnyer DM, Nicholls L, Verfaellie M. The role of VMPC in metamemorial judgments of content retrievability. *Journal of Cognitive Neuroscience*. 2005; 17:832–846. [PubMed: 15904549]
- Sharot T, Riccardi AM, Raio CM, Phelps EA. Neural mechanisms mediating optimism bias. *Nature*. 2007; 450:102–105. [PubMed: 17960136]
- Skinner EI, Fernandes MA. Neural correlates of recollection and familiarity: A review of neuroimaging and patient data. *Neuropsychologia*. 2007; 45:2163–2179. [PubMed: 17445844]
- Steele JD, Currie J, Lawrie SM, Reid I. Prefrontal cortical abnormality in major depressive disorder: A stereotactic meta-analysis. *Journal of Affective Disorders*. 2006; 101:1–11. [PubMed: 17174405]
- Stuss DT, Benson DF. Neuropsychological studies of the frontal lobes. *Psychological Bulletin*. 1984; 1:3–28. [PubMed: 6544432]
- Taylor SE, Brown JD. Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*. 1988; 103:193–210. [PubMed: 3283814]
- Tversky A, Kahneman D. Judgement under uncertainty: Heuristics and biases. *Science*. 1974; 185:1124–1131. [PubMed: 17835457]
- Uddin LQ, Iacoboni M, Lange C, Keenan JP. The self and social cognition: The role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences*. 2007; 11:153–157. [PubMed: 17300981]
- Vogeley K, Bussfeld P, Newen A, Hermann S, Happe F, Falkai P, et al. Mind reading: Neural mechanisms of theory of mind and self-perspective. *Neuroimage*. 2001; 14:170–181. [PubMed: 11525326]
- Volkow ND, Fowler JS, Wolf AP, Hitzemann R, Dewey S, Bendriem B, et al. Changes in brain glucose metabolism in cocaine dependence and withdrawal. *American Journal of Psychiatry*. 1991; 148:621–626. [PubMed: 2018164]
- Wheeler ME, Buckner RL. Functional-anatomical correlates of remembering and knowing. *Neuroimage*. 2004; 21:1337–1349. [PubMed: 15050559]

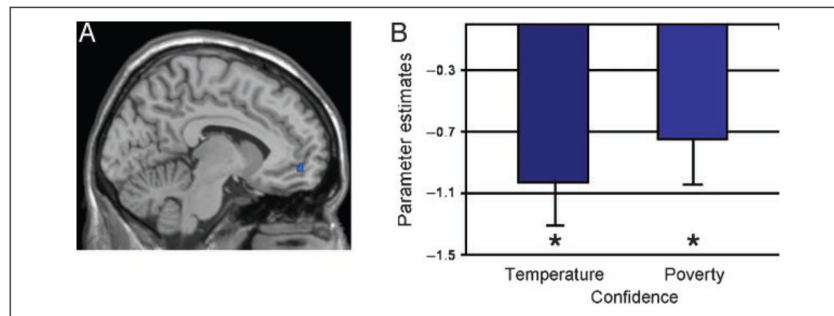


Figure 1. mPFC deactivation (peak BA 10, $x = -5$) associated with on-line self-evaluations of confidence. (A) Conjunction analysis of confidence estimates in relation to baseline. (B) Parameter estimates of mPFC activation in relation to baseline. *Parameter estimates significantly different than baseline.

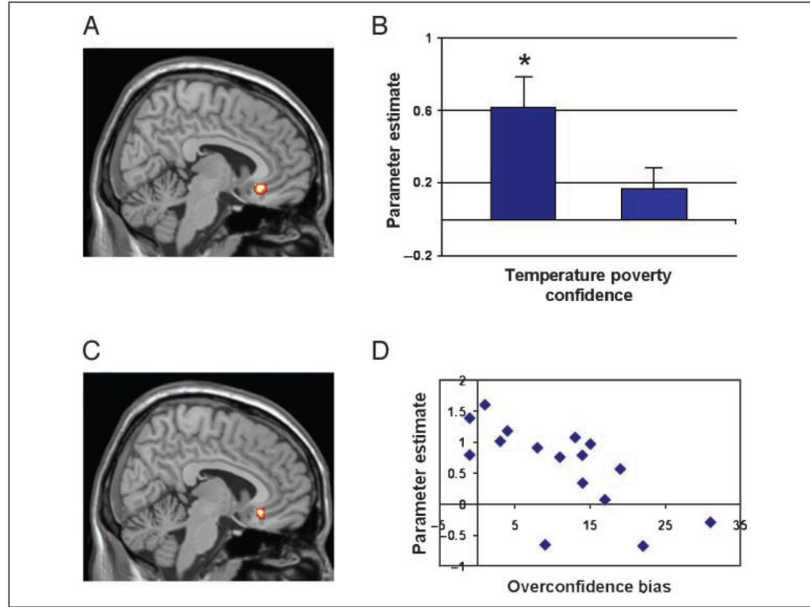


Figure 2. OFC activation (peak BA 11, $x = 7$) associated with overconfident self-evaluations. (A) Contrast between confidence estimates in the temperature condition and the poverty condition (collapsed across correctness of reasoning trial). (B) Parameter estimates of OFC activation for each confidence condition in relation to baseline. *Parameter estimates significantly different than baseline. (C) Regression analysis with magnitude of overconfident beliefs. (D) Parameter estimates in the OFC in relation to magnitude of overconfident beliefs for the temperature confidence estimate condition.

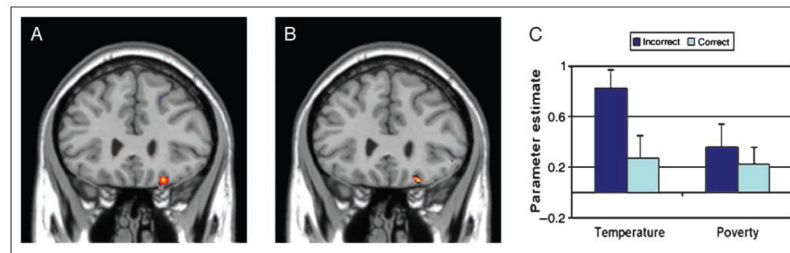


Figure 3.

OFC activation (peak BA 11, $y = 28$) that is down modulated by increasing levels of confidence. (A) Contrast between confidence estimates in the temperature condition and the poverty condition. (B) Parametric regressor that is negatively associated with confidence level in the temperature condition. (C) Parameter estimates of OFC activation for each confidence condition in relation to baseline. *Parameter estimates significantly different than baseline.

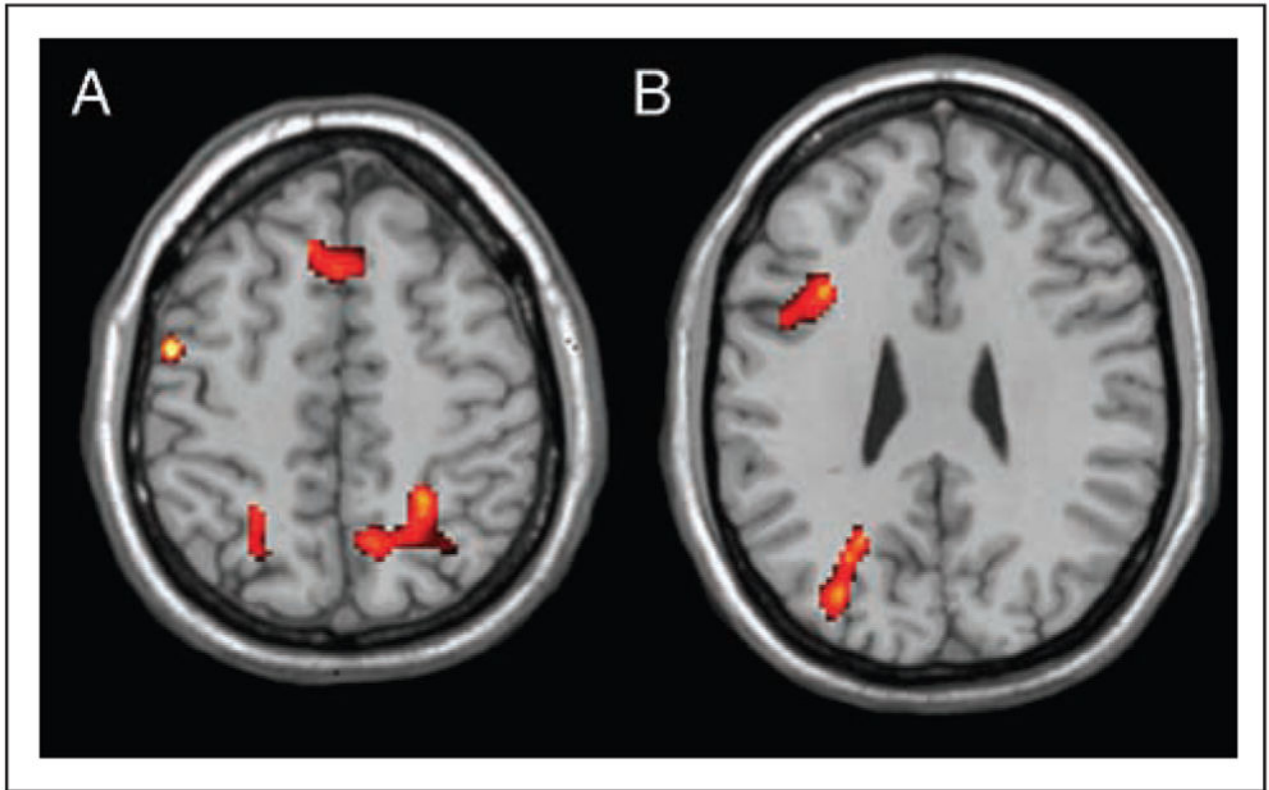


Figure 4. Left lateral prefrontal and parietal regions associated with reasoning in the condition associated with overconfident self-evaluations (temperature) compared with reasoning in the condition associated with accurate self-evaluations (poverty). (A) $z = 50$; (B) $z = 28$.

Group Activations and Deactivations Associated with the Conjunction of the Confidence Estimates across the Temperature and Poverty Conditions

Table 1

Region of Activation (Right/Left)	Brodmann's area	Coordinates			t
		x	y	z	
<i>Deactivation</i>					
mPFC (R/L)	10	-6	52	-6	4.16
<i>Activations</i>					
Superior frontal gyrus (R)	9	30	46	42	7.30
Middle frontal gyrus (R)	46				
Middle frontal gyrus (L)	9	-38	44	42	5.12
Superior frontal gyrus (L)	6	-20	-2	56	6.52
Superior frontal gyrus (R)	6	24	4	58	5.74
Supplementary motor area (L)	6	-4	10	54	5.19
Inferior parietal cortex (L)	40	-36	-36	42	5.80
Lingual gyrus (R/L)	17	-22	-76	-14	9.96

Thresholded at $p < .005$, uncorrected, $k = 15$ voxels. Approximate Brodmann's areas are shown from the Automated Anatomical Labeling map.

Table 2

Group Activations Associated with Overconfident Estimates in Comparison to Relatively Accurate Confident Estimates

Region of Activation (Right/Left)	Brodman's area	Coordinates			t
		x	y	z	
<i>Temperature > Poverty</i>					
Orbitofrontal cortex (R/L)	11	-6	26	-12	6.17
Orbitofrontal cortex (R)	11	12	50	-12	4.20
Orbitofrontal cortex (R)	11	20	30	-24	4.34
Middle temporal gyrus (L)	21	-68	-28	0	6.29
Calcarine (L)	17	-8	-88	20	4.31
Cuneus (L)	18	-8	-92	6	4.04
<i>Poverty > Temperature</i>					
Superior frontal gyrus (R)	6	14	14	-4	6.71
Inferior frontal cortex (L)	47	-40	36	-18	3.35
Supplementary motor area (L)	6	-6	-4	64	5.72
Precentral gyrus (R)	6	30	-12	66	5.57
Parietal cortex (L)	7	-36	-40	38	4.46
Parietal cortex (L)	7	-22	-56	46	4.26
Fusiform gyrus (L)	37	-38	-52	-18	3.34
Fusiform gyrus (L)	37	-26	-44	-18	3.72
Fusiform gyrus (R)	37	26	-50	-16	3.39
Supplementary motor area (R)	8	2	20	48	3.56
Lingual gyrus (L)	19	-28	-64	0	3.51
Occipital cortex (L)	37	-48	-68	-8	3.49
Parietal cortex (R)	7	26	-68	42	3.38
Precuneus (R)	5	12	-60	56	5.55

Thresholded at $p < .005$, corrected for areas significantly activated in the main effect of either condition, $k = 15$ voxels. Approximate Brodman's areas are shown from the Automated Anatomical Labeling map.

Table 3

Group Activations Associated with the Reasoning Tasks

Region of Activation (Right/Left)	Brodmann's area	Coordinates			t
		x	y	z	
<i>Temperature > Poverty</i>					
Inferior temporal gyrus (R)	20	58	-44	-26	4.06
Fusiform cortex (L)	37	-38	-50	-16	5.85
Operculum (L)	44	-32	16	28	4.26
Inferior frontal gyrus (L)	45	-34	32	14	3.66
Supplementary motor cortex (R)	8	4	22	54	4.31
Supplementary motor cortex (L)	6	-6	-4	64	5.06
Superior frontal gyrus (L)	8	-24	22	62	5.12
Superior frontal gyrus (R)	8	32	8	66	5.07
Superior frontal gyrus (L)	6	-26	2	68	5.00
Middle frontal gyrus (L)	9	-36	34	48	5.02
Middle frontal gyrus (R)	46	40	36	42	3.28
Precentral gyrus (L)	6/44	-50	-4	50	5.09
Precentral gyrus (R)	44	40	4	30	3.23
Postcentral gyrus (L)	48	-46	-6	18	3.74
Parietal cortex (R)	7	26	-64	38	6.78
Parietal cortex (L)	19	-28	-76	24	5.40
Occipital cortex (R)	19	36	-82	-4	6.50
<i>Poverty > Temperature</i>					
Middle temporal gyrus (R)	22	52	-46	18	5.26
Middle temporal gyrus (L)	21	-68	-28	2	4.78
Superior temporal gyrus (L)	22	-52	-22	12	3.96
Cingulate cortex (L/R)	23	0	-18	40	5.19
Cingulate cortex (R)	24	2	14	38	3.72
Medial frontal cortex (L)	9	-6	48	46	4.87
Medial frontal cortex (L)	10	-8	56	26	4.66
Ventromedial frontal cortex (R)	10	10	48	-8	4.53

Region of Activation (Right/Left)	Brodmann's area	Coordinates			t
		x	y	z	
Cuneus (L)	18	-8	-86	18	6.88
Calcarine (L)	19	-24	-62	8	4.58

Thresholded at $p < .005$, corrected for areas significantly activated in the main effect of either condition, $k = 15$ voxels. Approximate Brodmann's areas are shown from the Automated Anatomical Labeling map.