NeuroImage 43 (2008) 775–783

Contents lists available at ScienceDirect

NeuroImage

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

The Quadruple Process model approach to examining the neural underpinnings of prejudice

Jennifer S. Beer [a,*], Mirre Stallen [b,c], Michael V. Lombardo [d], Karen Gonsalkorale [g], William A. Cunningham [f], Jeffrey W. Sherman [e]

[a] Department of Psychology, #A8000 University of Texas at Austin, Austin, TX 78712, USA
[b] Erasmus University Rotterdam, Netherlands
[c] Radboud University Nijmegen, Netherlands
[d] University of Cambridge, UK
[e] University of California, Davis, USA
[f] Ohio State University, USA
[g] University of Sydney, Australia

ARTICLE INFO

Article history:
Received 3 December 2007
Revised 30 June 2008
Accepted 20 August 2008
Available online 6 September 2008

ABSTRACT

In order to investigate the systems underlying the automatic and controlled processes that support social attitudes, we conducted an fMRI study that combined an implicit measure of race attitudes with the Quadruple Process model (Quad model). A number of previous neural investigations have adopted the Implicit Association Test (IAT) to examine the automatic processes that contribute to social attitudes. Application of the Quad model builds on this previous research by permitting measures of distinct automatic and controlled processes that contribute to performance on the IAT. The present research found that prejudiced attitudes of ingroup favoritism were associated with amygdala, medial and right lateral orbitofrontal cortex. In contrast, prejudiced attitudes of outgroup negativity were associated with caudate and left lateral orbitofrontal cortex. Frontal regions found in previous neural research on the IAT, such as anterior cingulate, dorsolateral prefrontal cortex and operculum were associated with detecting appropriate responses in situations in which they conflict with automatic associations. Insula activity was associated with attitudes towards ingroup and outgroup members, as well as detecting appropriate behavior.

Introduction

The most complete understanding of neural systems underlying social attitudes must include systems underlying both automatic and controlled processes. For example, in the case of prejudice, a person may have an automatic tendency to judge an outgroup member in a negative manner or an ingroup member in a positive manner but control the expression of those automatic tendencies for a variety of reasons (Monteith et al., 1998; Sherman, in press). In other words, automatic associations may not be expressed because of controlled processing. In order to address the problematic behavioral measurement of automatic aspects of attitudes, a small amount of neural research interested in the neural systems underlying prejudice has drawn on the Implicit Association Test (IAT: Greenwald et al., 1998) (Chee et al., 2000; Cunningham et al., 2004; Knutson et al., 2006; 2007; Luo et al., 2006; Phelps et al., 2003; Phelps et al., 2000; Richeson et al., 2003). Although the IAT is one of the most popular behavioral measures of automatic (or implicit) attitudes, particularly for prejudice (Fazio and Olson, 2003), behavioral and modeling research suggests that IAT performance reflects both automatic and controlled processes (Conrey et al., 2005; Sherman et al., 2008). Previous fMRI studies involving the IAT are consistent with this view; significant activation has been found in regions associated with automatic processing, such as the amygdala (Cunningham et al., 2004; Phelps et al., (2000) although see Phelps et al., (2003) for evidence that amygdala is not critical for implicit racial bias) as well as controlled processing such as the dorsolateral prefrontal cortex (Richeson et al., 2003), middle frontal gyrus (Knutson et al., 2007), ventrolateral prefrontal cortex (Luo et al., 2006), and cingulate gyrus (Luo et al., 2006). The proposed research builds on this work by applying the Quadruple Process model (Quad model; Conrey et al., 2005; Sherman et al., 2008) to the interpretation of IAT performance in an fMRI environment, in order to examine the neural correlates of specific automatic and controlled processes that contribute to prejudice.

Automatic and controlled components of the IAT

The IAT measures attitudes by examining differential abilities to associate targets and attributes (e.g., targets of race: Black or White;

* Corresponding author. Fax: +1 512 471 5935.
E-mail address: jsbeer@ucdavis.edu (J.S. Beer).

1053-8119/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2008.08.033

attributes: Pleasant or Unpleasant). Participants are presented with stimuli and must categorize them using response keys that are associated with both a target and an attribute. In the Congruent condition of the race version of the task, the response keys represent pairings of targets and attributes that reflect negatively biased associations towards Black targets (e.g., Black/Unpleasant for one response key; White/Pleasant for the other response). In the Incongruent condition of this task, response keys represent pairings reflecting negative associations towards White targets (e.g., Black/Pleasant for one response key, White/Unpleasant for the other response key). Implicit bias is indicated by significantly slower response times in the Incongruent condition compared to the Congruent condition. In other words, if participants are slower to categorize stimuli when target and attribute pairings reflect negative Black/positive White associations in comparison to pairings that reflect positive Black/negative White associations, then they are considered to have negative implicit attitudes towards Black targets and/or positive implicit attitudes towards White targets.

Although the IAT is often used to measure automatic aspects of bias, performance of the task also recruits a variety of controlled processes (Conrey et al., 2005; Sherman et al., 2008). Consider the Stroop Task (Stroop, 1935), which is highly similar in structure to the IAT. A young child who knows colors but does not know how to read and a fully literate adult may make an equally small number of errors on the task. However, very different processes are at work for the adult and the child. On incongruent trials (e.g., the word "Blue" written in red ink), the adult must overcome a habit to read the word in order to name the color of the ink correctly. In contrast, the child has no habit to overcome; s/he simply responds to the color of the ink.

The same principle applies to the IAT (and many other implicit measures of attitudes), which has a Stroop-like structure of Congruent (e.g., pairing Black faces with negative words and White faces with positive words) and Incongruent (e.g., pairing Black faces with positive words and White faces with negative words) trials. The same behavioral outcome may reflect very different underlying processes. Whereas some people may exhibit weak bias because they successfully overcome their strong automatic evaluative associations, others may exhibit weak bias because they do not hold biased attitudes. The measure itself cannot distinguish between the two cases.

Previous neuroimaging research using the IAT

The joint contribution of automatic and controlled processes to IAT performance has been reflected in fMRI studies of this topic (for lesion studies using the IAT see Milne and Grafman (2001); Phelps et al., 2003). Previous neural research that combines the IAT and fMRI usually takes one of two approaches. In one approach, individual differences in IAT performance from outside the scanner are examined in relation to neural activity from a separate task in the scanner. Studies using this approach have focused on racial bias and have found significant amygdala activity (Cunningham et al., 2004; Phelps et al., 2000) and dorsolateral prefrontal cortex activity (Richeson et al., 2003) when viewing unfamiliar Black faces in comparison to White faces. These studies suggest that racial bias predicts the engagement of neural regions associated with both automatic processing (e.g., amygdala) and controlled processing (e.g., dorsolateral prefrontal cortex) when viewing outgroup faces in relation to ingroup faces.

In a second approach, neural activity is examined while participants perform the IAT in the scanner (attitudes toward objects in the natural world: Chee et al., 2000; moral issues: Luo et al., 2006; political issues: Knutson et al., 2006; gender and race: Knutson et al., 2007). As mentioned above, the behavioral measure of bias in an IAT paradigm rests on the discrepancy between reactions in the Incongruent and Congruent conditions. Therefore, fMRI studies of IAT performance typically compare neural activity in the Incongruent condition to the Congruent condition. These studies have found neural regions associated with controlled processes to be more active in Incongruent than Congruent conditions (middle frontal gyrus: Knutson et al., 2007; ventrolateral prefrontal cortex and anterior cingulate: Luo et al., 2006; left inferior frontal gyrus: Knutson et al., 2006; Chee et al., (2000) did not design their study to permit a direct comparison between the Incongruent and Congruent conditions). However, although the discrepancy in reaction time between the Incongruent and Congruent conditions is the basis for behaviorally measuring implicit bias, it is unlikely that neural activity associated with this comparison measures automatic components of attitudes in an fMRI environment. From an fMRI perspective, the comparison of the Incongruent and Congruent conditions represents the difference between neural systems that are engaged for a condition that may involve response competition and a condition with less response competition (rather than the comparison of the presence versus absence of an automatic association). In other words, the main contrast used in behavioral research to measure automatic bias translates into an analysis of controlled processes when conducted in the fMRI environment.

Therefore, the neural systems that support automatic attitudes in fMRI studies are often estimated by conducting other analyses within the IAT task structure. For example, implicit moral attitudes were examined by investigating the neural systems activated in relation to arousing moral target stimuli compared to non-arousing moral stimuli. Although arousal is equated with automaticity, it is possible that greater control is engaged in relation to arousing stimuli and, therefore, any results may reflect a combination of automatic and controlled processing. This analysis found significant activity in regions more typically associated with automatic associations for the arousing moral stimuli (amygdala, ventromedial prefrontal cortex: Luo et al., 2006). Another study examined implicit preference for political figures by regressing individual differences in explicit ratings of preference for the political figures on the Congruent condition map (in relation to a control condition). Although preferences were inferred as implicit, they were measured using explicit ratings, raising the possibility that both automatic and controlled processing contribute to the results. This analysis found significant activity in regions associated with automatic associations and regions associated with controlled processing (e.g., left superior frontal gyrus (BA 10), medial frontal gyrus (BA 11), right precentral gyrus (BA 6) and middle frontal gyrus (BA 8): Knutson et al., 2007). These studies illustrate the difficulty in using the IAT as a measure of automatic associations in an fMRI environment. Although automatic associations may contribute to arousal or explicit preferences, these measures may also reflect controlled processes.

In summary, previous fMRI studies of social attitudes using the IAT have found significant activation in neural systems associated with automatic processing and neural systems associated with controlled processing. One way of building on this research is to take an approach that permits researchers to relate this neural activity to specific automatic and controlled psychological processes. The present research illustrates one way to achieve this approach by applying the Quad model to the analysis of an IAT performed inside the scanner.

The Quad model

The Quad model was developed by Conrey et al. (2005; Sherman et al., 2008) to measure the joint contribution of automatic and controlled processes to performance on implicit measures of cognition. The Quad model is a multinomial model (Batchelder and Riefer, 1999) that measures the independent influences of four qualitatively different processes on implicit task performance by estimating a parameter value for each: automatic activation of an association with the stimulus (AC), the ability to detect an appropriate response (D), the success at overcoming automatically activated biased associations (OB), and the influence of any response bias that may guide overt responses when other guides to response are absent (G). The Activation parameter (AC) refers to the degree to which biased

associations are automatically activated when responding to a stimulus. All else being equal, the stronger the associations, the more likely they are to be activated and to influence behavior. The Detection parameter (D) reflects a relatively controlled process that detects appropriate and inappropriate responses. Sometimes, the activated associations conflict with the detected correct response. For example, on incompatible trials of the Stroop Task (Stroop, 1935) or incompatible trials of implicit attitude measures (e.g., pairing Black faces with positive words), automatic associations or habits conflict with detected correct responses. In such cases, the Quad model proposes that an Overcoming Automatically Activated Biased Associations process resolves the conflict. As such, the Overcoming Biased Associations parameter (OB) refers to self-regulatory efforts that prevent automatically activated associations from influencing behavior when they conflict with detected correct responses. Finally, the Guessing parameter (G) reflects general response tendencies that may occur when individuals have no associations that direct behavior, and they are unable to detect the appropriate response. Guessing can be random, but it may also reflect a systematic tendency to prefer a particular response. For example, incorrectly categorizing a target face stimulus as "unpleasant" in the IAT could be considered a socially undesirable response. To avoid that possibility, participants may adopt a conscious guessing strategy to respond with the positive rather than the negative key. Thus, guessing can be relatively automatic or controlled.

The structure of the Quad model is depicted as a processing tree in Fig. 1. In the tree, each path represents a likelihood. Processing parameters with lines leading to them are conditional upon all preceding parameters. For instance, Overcoming Biased Associations (OB) is conditional upon both Activation of Associations (AC) and Detection (D). If no automatic association exists, participants may still be able to detect an appropriate response (D) using information other than an automatic association, but OB cannot be calculated because there is no automatic association to overcome. Similarly, Guessing (G) is conditional upon the lack of Activation of Associations (1−AC) and the lack of Detection (1−D). The conditional relationships described by the model form a system of equations that predict the number of correct and incorrect responses in different conditions (e.g., compatible and incompatible trials). For example, a Black face stimulus in an incompatible block of a Black–White IAT will be assigned to the correct side of the screen with the probability: $AC\times D\times OB+(1-AC)\times D+(1-AC)\times(1-D)\times G$. This equation sums the three possible paths by which a correct answer can be returned in this case. The first part of the equation, $AC\times D\times OB$, is the likelihood that the association is activated and that the correct answer can be detected and that the association is overcome in favor of the detected response. The second part of the equation, $(1-AC)\times D$, is the likelihood that the association is not activated and that the correct response can be detected. Finally, $(1-AC)\times(1-D)\times G$, is the likelihood that the association is not activated and the correct answer cannot be detected and that the participant guesses by pressing the positive ("pleasant") key. Because the "pleasant" and "Black" categories share the same response key in the incompatible block, pressing the positive key in response to a Black face stimulus will return the correct answer. The respective equations for each item category (e.g., Black faces, White faces, positive words, and negative words in both compatible and incompatible blocks) are then used to predict the observed proportion of errors in a given data set. The model's predictions are then compared to the actual data to determine the model's ability to account for the data. A χ^2-estimate is computed for the difference between the predicted and observed errors. In order to best approximate the model to the data, the four parameter values are changed through maximum likelihood estimation until they produce a minimum possible value of the χ^2. The final parameter values that result from this process are interpreted as relative levels of the four processes. For a complete description of data analysis within the Quad model, see Conrey et al., (2005).

To date, the Quad model has been applied to and has been shown to accurately predict behavior on a variety of priming tasks, including semantic priming tasks (Gawronski and Bodenhausen, 2005; Sherman et al., 2008) and the weapon identification task (Conrey et al., 2005; Payne, 2001; Sherman et al., 2008). The model also has been applied extensively to the IAT; (Conrey et al., 2005; Greenwald et al., 1998; Sherman et al., 2008) and the Go/No-Go Association Task (GNAT; Nosek and Banaji, 2001; Gonsalkorale et al., in press; Sherman et al., 2008).

Validation of the Quad model

The viability of the Quad model depends on four critical elements: model fit (i.e., does the model adequately approximate behavioral data?), stochastic validity of the parameters (i.e., can the model's parameters be influenced independently?), construct validity of the parameters (i.e., do the parameters signify the processes claimed by the model?), and predictive validity of the parameters (i.e., do the parameters predict meaningful behaviors?). The Quad model has succeeded on all fronts. As described above, the model has

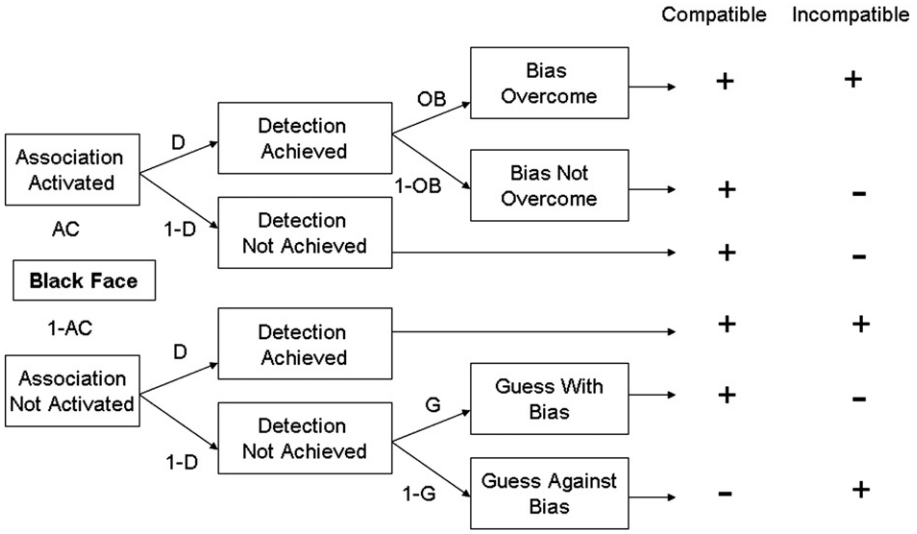

Fig. 1. The Quadruple Process model (Quad model). Each path represents a likelihood. Parameters with lines leading to them are conditional upon all preceding parameters. The table on the right side of the figure depicts correct (+) and incorrect (−) responses as a function of process pattern and trial type.

shown its ability to accurately predict performance on a variety of priming tasks, IATs, and the GNAT, demonstrating good model fit for these tasks (Conrey et al., 2005; Gonsalkorale et al., in press; Sherman et al., 2008).

Stochastic validity

The stochastic validity of the model has been established in a number of ways (Conrey et al., 2005; Sherman et al., 2008). For example, implementing a response deadline in an IAT designed to assess implicit attitudes about flowers and insects reduced Detection (D) and Overcoming Biased Associations (OB), but left Association Activation (AC) and Guessing (G) unaffected. Manipulating the base rate of left-hand versus right-hand responses in the same task affected Guessing (G), but none of the other three parameters (AC, D, OB). The expectation that one's performance on the weapon identification task would be observed by others decreased participants' ability to detect the appropriate response (D), but increased success at Overcoming Biased Associations (OB). These results indicate that the four parameters of the Quad model can vary independently, providing clear evidence for the stochastic validity of the model.

Construct validity

The construct validity of the model parameters also has been established by a number of findings (Conrey et al., 2005; Sherman et al., 2008). The fact that Detection (D) and Overcoming Biased Associations (OB) were reduced by a response deadline supports the claim that the two parameters reflect controlled processes that require cognitive capacity. In contrast, the finding that activation (AC) and Guessing (G) were unaffected by the response deadline is consistent with their depiction as relatively automatic processes that do not require significant cognitive capacity. The validity of OB as a measure of self-regulation was further established by demonstrations that it is impaired by alcohol consumption and decreases with age (Sherman et al., 2008). Extensive research has shown both alcohol use (e.g., Easdon and Vogel-Sprott, 2000) and aging (e.g., Hasher and Zacks, 1988) to be associated with impairments in self-regulation. The fact that altering the base rate of left-hand and right-hand responses influenced G corroborates the portrayal of that parameter as a general response bias.

Predictive validity

Two studies provide evidence for the predictive validity of the parameters. First, estimates of individual subjects' Association Activation (AC) parameters derived from an evaluative IAT were positively correlated with association-related reaction time impairment in the same task (Conrey et al., 2005). Thus, the higher the AC, the greater the association-based impairment in performance. At the same time, estimates of subjects' Overcoming Biased Associations (OB) parameters were negatively correlated with association-based reaction time impairment. Thus, the higher the OB, the better able were participants to avoid association-based impairments in performance. These findings also bolster the construct validities of the AC and OB parameters.

In another study (Gonsalkorale et al., in press), non-Muslim Caucasian participants interacted with an experimental confederate who appeared to be and was described as Muslim. Following the interaction, the confederate rated how much he liked the participants, while the participants completed a GNAT measuring implicit bias toward Muslims. The confederate's ratings of how much he liked the participants were predicted by an interaction between the AC and OB parameter estimates taken from the GNAT. Specifically, when participants had low AC estimates of negative associations with Muslims, their level of OB was unrelated to how much they were liked by the confederate. In contrast, participants with high AC estimates of negative associations with Muslims were liked to the extent that they had high OB parameter estimates. Thus, the ability to overcome automatic negative associations on the GNAT predicted the quality of the social interaction when those associations were strong.

In sum, the Quad model has shown its ability to accurately describe behavior on a variety of evaluative priming tasks, semantic priming tasks, IATs, and the GNAT. In addition, the stochastic and construct validities of the model's parameters have been supported by numerous findings. Finally, the predictive validity of the AC and OB parameters has been demonstrated.

Overview of the present study

The present study uses functional magnetic resonance imaging (fMRI) to examine the neural correlates of prejudice by applying the Quad model to a race IAT performed in an MRI scanner. This approach examines neural activity that is directly related to performing the IAT and generates individual difference measures of automatic and controlled processes that can be used as regressors on this neural activity. Additionally, the Quad model permits measurement of the automatic and controlled processing at a level of specificity not possible in previous studies. The relative nature of the IAT measure (i.e., difference score) conceals the different contributions of ingroup favoritism and outgroup hostility to performance. Unlike previous research that has correlated the relative preference for Black and White (Good minus Bad), the Quad model provides separate estimates of positive ingroup associations and negative outgroup associations, which will refine our understanding of the psychological nature of the activations found in previous IAT research (e.g., amygdala, medial frontal lobe, insula: Cunningham et al., 2004; Knutson et al., 2007; Phelps et al., 2000). In other words, this study will be the first to disentangle two distinct processes that contribute to prejudice — ingroup favoritism (positive associations) and outgroup negativity (negative associations). Further, application of the Quad model will also refine our understanding of the frontal lobe activations found in previous research on the IAT (e.g., Chee et al., 2000; Luo et al., 2006; Knutson et al., 2006; 2007; Richeson et al., 2003) by generating independent estimates of two distinct controlled processes (D, OB) that can be related to neural activity.

Materials and methods

Participants

Sixteen right-handed Caucasian participants (8 female; M=24.3 years, SD=4.6 years) were recruited in compliance with the human-subjects regulations of the University of California, Davis, and were compensated with *$10/h or course credits for their participation.

Behavioral paradigm

Participants completed a Black–White IAT designed to assess implicit preference for Whites over Blacks. Stimuli for the IAT consisted of 8 Pleasant and 8 Unpleasant pictures (from the International Affective Picture Set (IAPS); Lang et al., 1995) and 8 pictures of Black and 8 pictures of White faces (Jarvis, 2006; Minear and Park, 2004).

Four runs of six alternating Congruent and Incongruent blocks were completed (twenty-four blocks in total). Fig. 2 portrays the experimental paradigm. In the Congruent blocks, participants were instructed to respond to Black faces and Unpleasant pictures with a left-hand key and to White faces and Pleasant pictures with a right-hand key. In the Incongruent blocks, the response pairings were switched (i.e., Black faces and Pleasant pictures were paired with a right-hand key and White faces and Unpleasant pictures were paired with a left-hand key). During scanning, participants were holding a response box in each hand and responded by pressing with their left

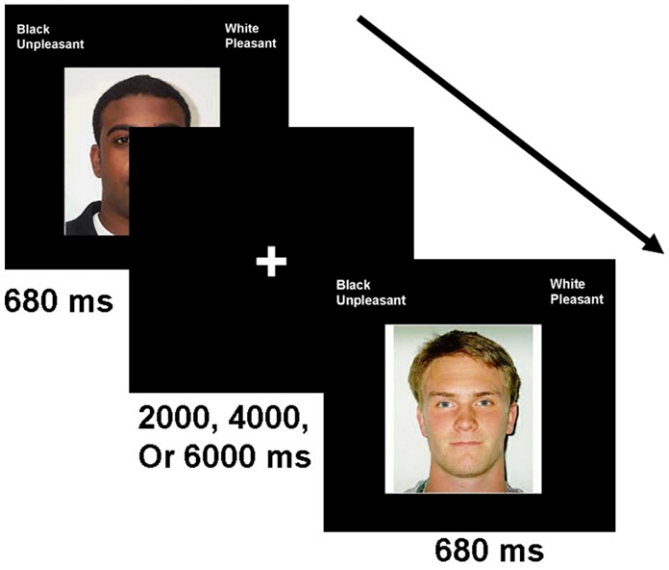

Fig. 2. Temporal layout of stimulus presentation in the Black–White IAT. Shown here is an example sequence of Congruent trials with the presentation of a Black face and a White face, separated by fixation crosses.

and right index fingers. Each of the blocks consisted of sixteen trials with four trials of each trial type: Unpleasant pictures, Pleasant pictures, White faces, and Black faces. Within a block, pictures were randomly selected without replacement, trials alternated between the presentation of faces and Pleasant/Unpleasant pictures, and each block started with the presentation of a face. At the beginning of each block, participants saw a screen indicating the responses and associated keys for that block. Participants were instructed to categorize the stimuli as quickly and accurately as possible. Each stimulus screen consisted of a picture in the center and a category-attribute pairing label in the left and right upper corner. The stimulus screen was presented for 680 ms (e.g., Cunningham et al., 2004), a time designed to increase errors on the task, which are used to estimate the Quad model parameters. The stimulus screen was followed by a screen with a fixation point. Participants were instructed to clear their minds when they saw the fixation point. The presentation of fixation points was jittered to optimize the estimation of the event-related BOLD response (50% were 2 s, 25% were 4 s, 25% were 6 s; Donaldson et al., 2001).

fMRI 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 (TR = 2000 ms, TE = 40 ms, FOV = 240, 64 × 64 matrix, voxel size 3.444 × 3.44 × 5 mm) with each volume consisting of 24 oblique axial slices which were tilted −15° from the AC–PC line. Both coplanar and high resolution T1-weighted images were also acquired from each subject so that functional data could be normalized to the Montreal Neurological Institute (MNI) atlas space. Head movement during acquisition was limited using foam inserts surrounding the head. Stimuli were projected onto a screen with a black background that was visible for the participants via a mirror attached to the head coil.

IAT and Quad model analysis

IAT bias was assessed by comparing error rates in Congruent and Incongruent trials. Five parameters were estimated for each participant: A measure of the automatic association between Black/Unpleasant (AC-Black/Unpleasant), a measure of the automatic association between White/Pleasant (AC-White/Pleasant), a measure of participants' ability to detect the appropriate response on the task (D), a measure of participants' ability to overcome automatic associations when they conflict with correct responses (OB), and a measure of a response bias to press the "pleasant" key (G). Parameter estimates of the Quad model were calculated using error rates on each trial type for each participant using the computation approach described by Conrey et al. (2005). The Quad model fit the behavioral data well ($\chi^2(1) = 1.93$, $p > .05$).

fMRI data analysis

SPM2 software (Wellcome Department of Cognitive Neurology, London) was used to pre-process the data. Images were corrected for differences in slice time acquisition and motion corrected using rigid-body transformation parameters. Anatomical images were coregistered to the mean functional image and spatially normalized to a T1 template. Templates were based on the Montreal Neurological Institute (MNI) stereotaxic space. The functional images were then normalized using those parameters and interpolated to 3 mm cubic voxels. Functional images were spatially smoothed with a Gaussian filter (8 mm. full width-half maximum). To remove drifts within sessions, a high-pass filter with a cutoff period of 128 s was applied.

A fixed-effects analysis was used to model event-related responses for each participant. Responses related to Congruent-faces, Congruent-pictures, Incongruent-faces, and Incongruent-pictures conditions were modeled with a canonical hemodynamic response function. A general linear model analysis then was used to create contrast images for each participant summarizing differences within and across the Incongruent and Congruent conditions as well as differences within and across Face and Picture conditions. These images were used to create group average SPM{t} maps that were thresholded at $p < 0.001$, 15 voxel minimum.

In order to examine the neural activity associated with the parameter values of the Quad model, the parameter estimates were regressed on relevant contrast maps using a region of interest (ROI) analysis to correct analysis for hypothesized neural regions. Most of the hypothesized neural regions were based on activations identified in previous neural investigations using the IAT. These regions included amygdala, insula, caudate, orbital, and medial and lateral portions of the frontal lobes (Chee et al., 2000; Cunningham et al., 2004; Knutson et al., 2006; 2007; Luo et al., 2006; Phelps et al., 2000; Richeson et al., 2003). These analyses were conducted using the AAL map to identify anatomical boundaries and an ROI toolbox designed for use with SPM ('MarsBar'; Brett et al., 2002) to correct for these regions. In order to examine the neural correlates of the AC-Black/Unpleasant, AC-White/Pleasant, D and OB parameters, individual parameter values were entered as regressors on relevant contrast maps, which were masked for regions of interest identified by previous research. AC-Black/Unpleasant parameter values were regressed on the Black faces > White faces contrast, AC-White/Pleasant parameter values were regressed on the White faces > Black faces contrast, D parameter values were regressed the Incongruent > Congruent contrast and OB parameter values were regressed on the Incongruent-faces > Congruent-faces map. The study focuses on (a) the automatic associations of ingroups and outgroups and (b) the controlled processes of detecting an appropriate response when it may conflict with an automatic association, and overcoming biased associations. The G parameter is not regressed on the neural maps because it is impossible for us to disentangle whether it represents a bias to respond with the dominant hand or a bias toward using the "positive" response. This ambiguity makes it difficult to interpret any possible neural findings and, in the case of a bias to use the right hand, is less interesting for understanding the neural systems associated with prejudice. The Quad model analyses were thresholded at $p > 0.01$, 15 voxel minimum.

Results

Behavioral results

In the Quad model, behavioral data analysis focuses on error rates rather than reaction times. The 680 ms response deadline used to increase errors constrains reaction times and, therefore, the typical race bias effect is examined by contrasting error rates in the Incongruent condition with error rates in the Congruent condition. Participants made significantly more errors in the Incongruent than the Congruent condition ($F(1,15)=21.42$, $p<.05$), replicating the typical race bias IAT effect. The overall error rate was 10%; see Table 1 for mean parameter estimates.

Imaging results

The neuroimaging results are presented in Table 2. First, we report significant neural activity in relation to the comparison between the Incongruent and Congruent condition. This is the contrast typically reported in previous fMRI research on IAT performance. Second, we report significant neural correlates of each of the parameter estimates from the Quad model.

The IAT effect: incongruent compared to Congruent

The contrast between the Incongruent and Congruent condition showed significant activation in the left anterior cingulate gyrus (BA 11/25, Fig. 3A), right operculum (BA 44), right lateral frontal cortex (BA 44), and precuneus (left BA 30, right BA 23).

Automatic associations

The AC-Black/Unpleasant and AC-White/Pleasant parameters estimate automatic associations for Black and White faces, respectively. Therefore, each of these parameters was regressed on the relevant contrast between Black and White faces. The AC-Black/Unpleasant parameter was significantly associated with activation in bilateral insula, the right caudate, and the left lateral orbitofrontal cortex (BA47, see Fig. 3C) for the contrast between Black faces and White faces. The AC-White/Pleasant parameter was significantly associated with activations in the right insula, right amygdala, medial orbitofrontal cortex (BA11, Fig. 3D), and right lateral orbitofrontal cortex (BA47) for the contrast between White faces and Black faces.

Detection of appropriate responses

Parameter D estimates the ability of participants to detect a correct response for the task which is particularly challenging in the Incongruent conditions of the IAT. In order to understand the neural regions that support the ability to detect a correct response when it potentially conflicts with an automatic association, parameter D was regressed on the contrast between the Incongruent and Congruent conditions. Parameter D was associated with significant activation in the right insula, bilateral anterior cingulate (BA32, Fig. 3B), right lateral frontal cortex (BA 46) and the operculum (BA 44) for the contrast between the Incongruent and Congruent conditions.

Table 1
Parameter estimates for the Black–White IAT

Parameter	Estimate
AC-Black-Unpleasant	0.03
AC-White-Pleasant	0.11
OB	0.21
D	0.87
G	0.63

Note. AC = Association Activation, OB = Overcoming Biased Associations, D = Detection, G = Guessing. For the G parameter, the value of .5 indicates random guessing.

Table 2
Significant areas of activation associated with the IAT effect (Incongruent versus Congruent) and the parameter estimates of the Quad model

Region of activation	Left/right Brodmann	*x*	*y*	*z*	*t*-score
Incongruent>Congruent					
Anterior cingulate gyrus	L BA11/25	−4	26	−8	5.51
Dorsolateral prefrontal cortex	R BA44	38	10	40	4.36
Operculum	R BA44	50	8	26	4.85
Precuneus	L BA30	−6	−54	20	4.29
	R BA23	4	−54	24	3.99
AC-Black-Unpleasant					
Insula	R BA47/48	32	24	−2	3.81
	L BA48	−32	−14	16	3.44
Caudate	R	14	16	16	3.5
Lateral orbitofrontal cortex	L BA47	−32	22	−18	3.47
AC-White-Pleasant					
Insula	R BA48	34	14	−16	4.14
Amygdala	R	30	0	−14	4.71
Medial orbitofrontal cortex	L BA 11	−14	26	−18	4.16
	L BA11	−14	46	−18	3.59
	R BA11	16	30	−20	3.37
Lateral orbitofrontal cortex	R BA47	36	26	−18	3.43
D					
Insula	R BA48	34	20	−16	4.71
Anterior cingulate gyrus	L BA32	−8	32	24	3.18
	R BA32	8	40	14	4.19
	R BA32	8	34	28	3.87
Dorsolateral prefrontal cortex	R BA46	30	58	28	4.84
	R BA46	32	56	14	4.6
	R BA46	34	48	14	3.67
Operculum	R BA44	48	8	18	3.2

Note. AC = Association Activation, OB = Overcoming Bias, D = Detection, G = Guessing, R = right, L = left, BA = Brodmann's area.

Overcoming Biased Associations

Whereas the D parameter is important for detecting appropriate responses, the OB parameter is associated with the ability to overcome the expression of an automatic association when it conflicts with the appropriate response. Therefore, OB is conditional upon both the presence of Association Activation (AC parameters) and Detection (D) in the Face condition. No significant activations were found when OB was regressed on the contrast between the Incongruent-face and Congruent-face condition.

Discussion

The present research investigated the neural systems underlying prejudice by applying the Quad model (Conrey et al., 2005; Sherman et al., 2008) to analyze IAT performance (Greenwald et al., 1998) in an fMRI environment. The addition of the Quad model extended previous investigations of the neural systems underlying social attitudes by examining neural systems in relation to specific automatic and controlled processes that contribute to implicit bias. The present research found that insula activity was associated with automatic negative associations with outgroup members (AC-Black/Unpleasant), positive associations with ingroup members (AC-White/Pleasant), and detection of appropriate behavior when race targets are paired with an incongruent valence (D). The pervasive insula activity is consistent with previous research associating insula with individual differences in non-racial prejudice (Krendl et. al., 2006), as well as general arousal and emotional processing (Britton et al., 2006).

The Quad model analysis built on previous research in several ways. First, the Quad model analysis permitted the examination of two kinds of prejudice: ingroup bias and outgroup bias. Significantly different neural activation was found for automatic associations depending on whether the target was outgroup or ingroup members. Negative associations with outgroup members were related to

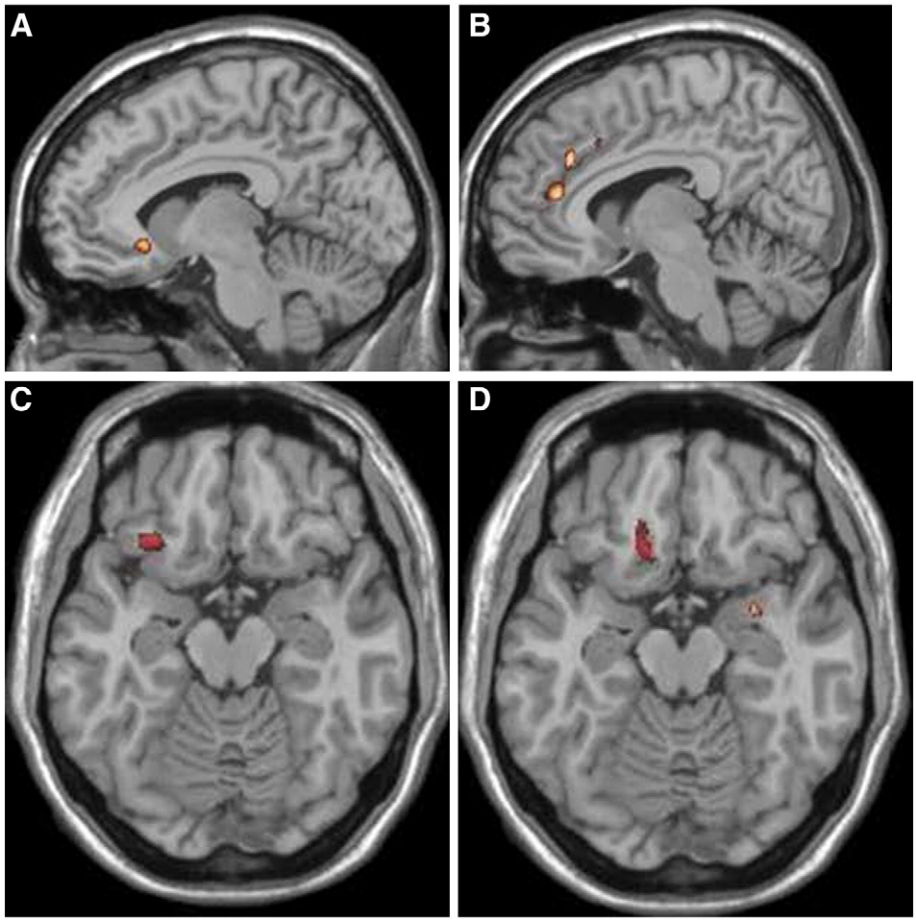

Fig. 3. Activation maps for brain areas. (A) IAT effect, left ACC ($x=-4$), whole brain-analysis for Incongruent>Congruent; (B) Right dACC ($x=8$), parameter D regressed on Incongruent>Congruent; (C) lateral OFC ($z=-16$), AC/Black-Unpleasant parameter regressed on Black>White; (D) Medial left OFC and right amygdala ($z=-16$), AC/White-Pleasant parameter regressed on White>Black.

caudate and left lateral orbitofrontal cortex activity, whereas positive associations with ingroup members were related to amygdala, medial orbitofrontal, and right lateral orbitofrontal activity. Additionally, the Quad model analysis built on previous research that examined controlled processes in prejudice through the comparison of subliminal and supraliminal stimuli (Cunningham et al., 2004). The previous research attempted to create conditions under which participants did not (i.e., subliminal) or did (i.e., supraliminal) have the ability to control a response. In contrast, the current research was able to directly measure the degree to which participants engaged in controlled processing. In the present study, the frontal regions associated with control over prejudice in previous research were significantly associated with participants' ability to detect appropriate behavior in a condition where it conflicts with automatic associations, but were not significantly associated with suppressing the expression of the automatic association (OB).

Automatic processing

Previous fMRI research that directly measured the IAT used a variety of approaches to examine the automatic associations underlying social attitudes, and found diverse neural activity including orbitofrontal cortex and amygdala activity (Cunningham et al., 2004; Luo et al., 2006; Phelps et al., 2000). The present research replicated and extended these results by associating these regions with independent estimates of positive ingroup and negative outgroup automatic associations. These findings suggest that the Quad model is particularly useful for understanding neural activity that reflects bias associated two kinds of prejudice: outgroup negativity and ingroup favoritism.

Negative associations with outgroup members were related to left lateral orbitofrontal cortex activity, whereas positive associations with ingroup members were related to amygdala, medial orbitofrontal cortex, and right lateral orbitofrontal cortex activity. The orbitofrontal cortex has been associated with processing both positive and negative information. It has been suggested that the medial orbitofrontal cortex may be more strongly associated with processing positive information, and that lateral orbitofrontal cortex may be more strongly associated with processing negative information (e.g., Kringelbach and Rolls, 2004). The medial and lateral distinction is somewhat supported by the findings in the present research; medial orbitofrontal cortex activity was only found for the positive association with ingroup members.

Although previous research has often been interpreted to suggest that amygdala activity indicates a threat response in relation to outgroup stimuli presented in the scanner (e.g. Cunningham et al., 2004; Phelps et al., 2000), the present research found amygdala activity in relation to implicit bias favoring ingroup members. These findings are consistent with more recent theories associating amygdala activity with environmentally significant stimuli (e.g., Whalen, 1998; Cunningham et al., 2008) that are rewarding (e.g., Baxter and Murray, 2002). Research has found amygdala activation in response to ingroup faces in conditions in which the ingroup is likely to be salient, such as a minimal group setting (Van Bavel et al., in

press), and in individuation judgments of ingroup members (Wheeler and Fiske, 2005). Consistent with this hypothesis, our sample held more of an ingroup bias (e.g., AC-White/Pleasant=.11) than an outgroup bias (AC-Black-Unpleasant=.03) suggesting a greater affective response to component of prejudice associated with White than Black. Finally, research suggests that interactions between amygdala and orbitofrontal cortex support responses in situations in which reward or threat outcomes are certain (Izquierdo and Murray, 2004; Schoenbaum, 2004). Therefore, the co-activation between amygdala and orbitofrontal cortex in relation to ingroup favoritism may reflect greater certainty that their response will be rewarded when responding to ingroup faces in comparison to outgroup faces.

Negative associations with outgroup members (AC-Black/Unpleasant) were also significantly associated with caudate activity, a region that has been found in previous studies of implicit attitudes when the Incongruent and Congruent conditions were compared (Luo et al., 2006). Previous research suggests that the caudate is involved in programming and initiating emotion-induced behavior, either for positive or negative emotions (Wager et al., 2003). This research suggests that the presentation of Black faces elicits a negative reaction, either because of prejudice or a fear of responding in a biased fashion, and the caudate is recruited to prepare a response.

Controlled processing

The present research also built on previous findings by associating neural activity with specific kinds of controlled processes that contribute to prejudiced responses. Control over implicit attitudes has previously been associated with dorsolateral prefrontal cortex, anterior cingulate, and operculum activity (the Incongruent condition of an IAT task: Chee et al., 2000; Luo et al., 2006; viewing outgroup faces: Cunningham et al., 2004; Richeson et al., 2003). The present study found significant activation in these same regions when participants detected appropriate responses (D) in conditions in which they conflict with automatic associations. The involvement of the ACC in the detection process is consistent with previous research that has found ACC activation in relation to the detection of conflict between possible responses in studies of implicit prejudice (e.g., Amodio et al., 2008; Richeson et al., 2003) and non-social tasks (e.g., Botvinick et al., 2004). Naturally, a necessary precondition for detecting conflict between responses is to detect what response is required. As such, the detection process described by the Quad model likely feeds into conflict detection processes. No significant activation was found in relation to inhibiting the expression of automatic bias (OB). This may have been due to the fact that OB is estimated from only a subset of the trials used to estimate neural activity in the Incongruent-faces>Congruent-faces contrast. Specifically, the structure of the model determines that OB is estimated only from trials on which both AC and D occur. The AC parameters for Black (.03) and White (.11) faces indicate that associations were activated in only a small subset of the trials, so there may have been too few instances to robustly estimate neural activity in relation to OB.

Conclusion

The present research exemplifies the new insights into the neural systems underlying social attitudes that can be gained from a combination of the Quad model, the IAT and neuroimaging. This approach permits investigators to more precisely identify neural activity in relation to automatic and controlled processes that support social attitudes and their expression. The Quad model proved helpful for specifically identifying neural systems that support separate ingroup and outgroup automatic associations, detecting appropriate responses that occur when there is conflict with automatic associations, and inhibition of automatic associations. The examination of these processes across domains will be helpful for understanding the distinction between automatic and controlled processing and distinctions among attitude domains.

References

Amodio, D.M., Devine, P.D., Harmon-Jones, E., 2008. Individual differences in the regulation of race bias: the role of conflict detection and neural signs for control. J. Pers. Soc. Psychol. 94, 60–74.

Batchelder, W.H., Riefer, D.M., 1999. Theoretical and empirical review of multinomial process tree modeling. Psychon. Bull. Rev. 6, 57–86.

Baxter, M.G., Murray, E.A., 2002. The amygdala and reward. Nat. Rev., Neurosci 3, 563–573.

Botvinick, M.M., Cohen, J.D., Carter, C.S., 2004. Conflict monitoring and anterior cingulate cortex: an update. Trends Cogn. Sci. 8, 539–546.

Brett, M., Anton, J.L., Valabregue, R., Poline, J.B., 2002. Region of interest analysis using an SPM toolbox. In Proc. 8th HBM, Sendai, Japan.

Britton, J.C., Taylor, S.F., Sudheimer, K.D., Liberzon, I., 2006. Facial expressions and complex IAPS pictures: common and differential networks. NeuroImage 31, 906–919.

Chee, M.W., Sriram, N., Soon, C.S., Lee, K.M., 2000. Dorsolateral prefrontal cortex and the implicit association of concepts and attributes. NeuroReport 11, 135–140.

Conrey, F.R., Sherman, J.W., Gawronski, B., Hugenberg, K., Groom, C.J., 2005. Separating multiple processes in implicit social cognition: the quad model of implicit task performance. J. Pers. Soc. Psychol 89, 469–487.

Cunningham, W.A., Johnson, M.K., Raye, C.L., Chris Gatenby, J., Gore, J.C., Banaji, M.R., 2004. Separable neural components in the processing of black and white faces. Psychol. Sci 15, 806–813.

Cunningham, W.A., Van Bavel, J.J., Johnsen, J.R., 2008. Affective flexibility: evaluative processing goals shape amygdala activity. Psychol. Sci. 19, 152–160.

Donaldson, D.I., Petersen, S.E., Ollinger, J.M., Buckner, R.L., 2001. Dissociating state and item components of recognition memory using fMRI. NeuroImage 13, 129–142.

Easdon, C.M., Vogel-Sprott, M., 2000. Alcohol and behavioral control: impaired response inhibition and flexibility in social drinkers. Exp. Clin. Psychopharmacol. 8, 387–394.

Fazio, R.H., Olson, M.A., 2003. Implicit measures in social cognition research: their meaning and use. Annu. Rev. Psychol 54, 297–327.

Gawronski, B., Bodenhausen, G.V., 2005. Accessibility effects on implicit social cognition: the role of knowledge activation and retrieval experiences. J. Pers. Soc. Psychol. 89, 672–685.

Gonsalkorale, K., von Hippel, W., and Sherman, J.W., in press. Bias and regulation of bias in intergroup interactions: implicit attitudes toward Muslims and interaction quality. J. Exp. Soc. Psychol.

Greenwald, A.G., McGhee, D.E., Schwartz, J.L., 1998. Measuring individual differences in implicit cognition: the implicit association test. J. Pers. Soc. Psychol 6, 1464–1480.

Hasher, L., Zacks, R.T., 1988. Working memory, comprehension, and aging: a review and a new view. In: Bower, G.H. (Ed.), The Psychology of Learning and Motivation: Advances in Research and Theory, Vol. 22. Academic Press, San Diego, pp. 193–225.

Izquierdo, A.D., Murray, E.A., 2004. Combined unilateral lesions of the amygdala and orbital prefrontal cortex impair affective processing in rhesus monkeys. J. Neurophysiol 91, 2023–2039.

Jarvis, W.B.G., 2006. DirectRT v2006 [computer software]. New York: Empirisoft Corporation.

Knutson, K.M., Wood, J.N., Spampinato, M.V., Grafman, J., 2006. Politics on the brain: an fMRI investigation. Soc. Neurosci 1, 25–40.

Knutson, K.M., Mah, L., Manly, C.F., Grafman, J., 2007. Neural correlates of automatic beliefs about gender and race. Hum. Brain Mapp. 28.

Krendl, A.C., Macrae, C.N., Kelley, W.M., Fugelsang, J.A., Heatheron, T.F., 2006. The good, the bad and the ugly: an fMRI investigation of the functional anatomic correlates of stigma. Soc. Neurosci. 1, 5–15.

Kringelbach, M.L., Rolls, E.T., 2004. The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. Prog. Neurobiol. 72, 341–372.

Lang, P.J., Bradley, M.M., Cuthbert, B.N., 1995. The international affective picture system (IAPS): Photographic slides. University of Florida: The Center for Research in Psychophysiology.

Luo, Q., Nakic, M., Wheatley, T., Richell, R., Martin, A., Blair, R.J., 2006. The neural basis of implicit moral attitude—an IAT study using event-related fMRI. NeuroImage 30, 1449–1457.

Milne, E., Grafman, J., 2001. Ventromedial prefrontal cortex lesions in humans eliminate implicit gender stereotyping. J. Neurosci 21, 1–6.

Minear, M., Park, D.C., 2004. A lifespan database of adult facial stimuli. Behav. Res. Meth. Instrum. Comput. 36 (4), 630–633.

Monteith, M.J., Sherman, J.W., Devine, P.G., 1998. Suppression as a stereotype control strategy. Personal. Soc. Psychol. Rev. 2, 63–82.

Nosek, B.A., Banaji, M.R., 2001. The go/no-go association task. Social Cogn. 19, 625–666.

Payne, B.K., 2001. Prejudice and perception: the role of automatic and controlled processes in misperceiving a weapon. J. Pers. Soc. Psychol. 81, 181–192.

Phelps, E.A., O'Connor, K.J., Cunningham, W.A., Funayama, E.S., Gatenby, J.C., Gore, J.C., Banaji, M.R., 2000. Performance on indirect measures of race evaluation predicts amygdala activation. J. Cogn. Neurosci. 12, 729–738.

Phelps, E.A., Cannistraci, C.J., Cunningham, W.A., 2003. Intact performance on an indirect measure of race bias following amygdala damage. Neuropsychologia 41, 203–208.

Richeson, J.A., Baird, A.A., Gordon, H.L., Heatherton, T.F., Wyland, C.L., Trawalter, S., Shelton, J.N., 2003. An fMRI investigation of the impact of interracial contact on executive function. Nat. Neurosci. 6, 1323–1328.
Schoenbaum, G., 2004. Affect, action, and ambiguity and the amygdala-orbitofrontal circuit. J. Neurophysiol. 91, 1938–1939.
Sherman, J.W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T.J., Groom, C.J., 2008. The self-regulation of automatic associations and behavioral impulses. Psychol. Rev. 115, 314–335.
Sherman, J.W., in press. Controlled influences on implicit measures: confronting the myth of process-purity and taming the cognitive monster. In: Petty, R.E., Fazio, R.H., Briñol P. (Eds.). Attitudes: insights from the new wave of implicit measures. Hillsdale, NJ: Erlbaum.
Stroop, J.R., 1935. Studies on the interference in serial verbal reactions. J. Exp. Psychol. 59, 239–245.
Van Bavel, J.J., Packer, D.J., Cunningham, W.A., in press. The Neural Substrates of In-Group Bias: A Functional Magnetic Resonance Imaging Investigation. Psychol. Sci.
Wager, T.D., Phan, K.L., Liberzon, I., Taylor, S.F., 2003. Valence, gender, and lateralization of functional brain anatomy in emotion: a meta-analysis of findings from neuroimaging. NeuroImage 19, 513–531.
Whalen, P.J., 1998. Fear, vigilance, and ambiguity: initial neuroimaging studies of the human amygdala. Curr. Dir. Psychol. Sci. 7, 177–188.
Wheeler, M.E., Fiske, S.T., 2005. Controlling racial prejudice: social-cognitive goals affect amygdala and stereotype activation. Psychol. Sci. 16, 56–63.