doi:10.1093/scan/nss063 SCAN (2013) 8, 750–755

Is race erased? Decoding race from patterns of neural activity when skin color is not diagnostic of group boundaries

Kyle G. Ratner,[1] Christian Kaul,[1,2] and Jay J. Van Bavel[1]
[1]Department of Psychology and [2]Center for Neural Science, New York University, New York, NY, USA

Several theories suggest that people do not represent race when it does not signify group boundaries. However, race is often associated with visually salient differences in skin tone and facial features. In this study, we investigated whether race could be decoded from distributed patterns of neural activity in the fusiform gyri and early visual cortex when visual features that often covary with race were orthogonal to group membership. To this end, we used multivariate pattern analysis to examine an fMRI dataset that was collected while participants assigned to mixed-race groups categorized own-race and other-race faces as belonging to their newly assigned group. Whereas conventional univariate analyses provided no evidence of race-based responses in the fusiform gyri or early visual cortex, multivariate pattern analysis suggested that race was represented within these regions. Moreover, race was represented in the fusiform gyri to a greater extent than early visual cortex, suggesting that the fusiform gyri results do not merely reflect low-level perceptual information (e.g. color, contrast) from early visual cortex. These findings indicate that patterns of activation within specific regions of the visual cortex may represent race even when overall activation in these regions is not driven by racial information.

Keywords: race; multivariate pattern analysis; fusiform gyrus; face network; fMRI

In perhaps his most famous speech, Martin Luther King Jr (1963) said, "I have a dream that my four children will one day live in a nation where they will not be judged by the color of their skin, but by the content of their character." Almost five decades after King spoke these words, America is a more integrated society, and social identities have become increasingly decoupled from the racial and ethnic cues that have historically defined them. Living in a pluralistic society often requires that people learn intergroup affiliations without relying on racial cues as a guide. As a consequence, people might not encode the race of other people's faces when it is not indicative of group boundaries (Sidanius and Pratto, 1999; Kurzban *et al.*, 2001; Cosmides *et al.*, 2003; Hehman *et al.*, 2010). However, racial differences are often associated with physiognomic markers, such as skin tone and facial features, and people can differentiate the race of faces within several hundred milliseconds (Caldara *et al.*, 2003; Ito and Urland, 2003). This study examines whether the visual system represents race when perceptual indicators of race are irrelevant to group membership.

Several neuroimaging studies have recently investigated the representation of race—a visually and socially salient social category—in the face processing network (see Macrae and Quadflieg, 2010 for a recent review). Although the neural correlates of face perception are widely distributed (Ishai *et al.*, 1999), a sub-region of the fusiform gyrus (FG), located on the ventral surfaces of the temporal lobe, plays a central role in the processing of faces (Sergent *et al.*, 1992; Puce *et al.*, 1995; Kanwisher *et al.*, 1997). Several functional magnetic resonance imaging (fMRI) studies have shown that race modulates neural activity in the FG. They specifically find increased activity in the FG in response to own-race *vs* other-race faces (Golby *et al.*, 2001; Lieberman *et al.*, 2005; Kim *et al.*, 2006).

Recent studies have questioned whether this reported racial bias in the FG response reflects factors related to race per se (e.g. expertise with own-*race* faces) or group membership more generally (e.g. identification with own-*group* faces). Specifically, Van Bavel *et al.* (2008, 2011) conducted a series of experiments in which White participants were assigned to one of two novel mixed-race groups and responded to Black and White faces from their in-group and an out-group. Making race orthogonal to group membership permitted an independent comparison of the effects of race (Black *vs* White) and group membership (in-group *vs* out-group). The faces in each group were counterbalanced across participants to ensure that any effects of group membership were due to group distinctions rather than exogenous stimulus properties. In both studies, greater activity to in-group *vs* out-group faces was found in the fusiform gyri (Van Bavel *et al.*, 2008), and, more specifically, the fusiform face area (Van Bavel *et al.*, 2011). Moreover, there was no main effect of race on the FG activity and the effect of group membership was not moderated by race. In addition, Van Bavel *et al.* (2011) found that the degree to which FG activity was greater to in-group *vs* out-group faces was correlated with better memory for in-group *vs* out-group faces. A series of behavioral follow-up studies found a similar pattern of results: participants showed preferences for in-group members on an implicit measure of evaluation (Van Bavel and Cunningham, 2009) and superior recognition memory for in-group faces (Van Bavel *et al.*, 2012), regardless of race.

Previous research examining activity in the FG in response to race and group membership used standard univariate fMRI analysis techniques (Van Bavel *et al.*, 2008, 2011). These univariate procedures average across the blood oxygen level dependent (BOLD) response recorded from a set of contiguous voxels in a particular brain area and test whether the resulting estimate is different between two or more stimuli or tasks (Friston *et al.*, 1995). Although univariate procedures are currently the conventional approach for analyzing fMRI data, there has been a growing interest in the neuroimaging community in using multivariate techniques, such as multivariate pattern

Received 11 November 2011; Accepted 24 May 2012
Advance Access publication 1 June 2012

We thank William Cunningham for generously funding the data collection phase of this research, and David Amodio, Tobias Brosch, Sharon David, Alumit Ishai, Dominic Packer, Chris Said, Jillian Swencionis, Jenny Xiao, Sophie Wharton and members of the *NYU Social Perception and Evaluation Lab* (@vanbavellab) for helping with various aspects of this research. We also thank Matthew Lieberman and two anonymous reviewers for thoughtful feedback on this manuscript. This research was supported by a National Science Foundation Graduate Research Fellowship to K.G.R., a Feodor-Lynen-Award from the Alexander von Humboldt Foundation to C.K. and a Social Sciences and Humanities Research Council of Canada Award to J.J.V.B. Data were collected at the Queen's University MRI facility. The first and second authors contributed equally to this manuscript.

Correspondence should be addressed to Jay J. Van Bavel, Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA. E-mail: jay.vanbavel@nyu.edu

analysis (Haynes and Rees, 2006; Norman *et al.*, 2006; Kriegeskorte *et al.*, 2006; Mur *et al.*, 2009).

Multivariate pattern analysis (MVPA) has recently been used successfully in a handful of studies to examine the representation of social categories (Chiu *et al.*, 2011; Kaul *et al.*, 2011; Natu *et al.*, 2010). Unlike traditional univariate analysis procedures, MVPA uses pattern classification algorithms to map categories of stimuli or psychological states to brain activity. In a typical experiment, a portion of the data is used to train classifiers to detect patterns of voxels that are responsive to specific conditions or categories of stimuli (e.g. Black and White faces). Then, the ability of the pattern of voxels that comprises each classifier to decode the remaining independent data is used to infer whether the conditions of interest are represented by different patterns of brain activity. Thus, whereas univariate analyses test differences between conditions at each individual voxel or the average of the voxels within a particular region, MVPA tests differences in *patterns* of voxels. In other words, MVPA allows investigators to examine whether different neural patterns of activation go undetected by traditional univariate analysis when two conditions produce the same mean-level of activation, but activate different voxels within a certain region of interest (ROI).[1]

As discussed earlier, univariate analyses have shown that group membership can influence the processing of faces in the absence of salient perceptual intergroup cues. In this context, group membership, and not race, appears to guide neural activity (Van Bavel *et al.*, 2008, 2011) and social behavior, including evaluation and recognition memory of faces (Van Bavel and Cunningham, 2009; Van Bavel *et al.*, 2012). These investigators also observed that the typical finding of greater fusiform activity to own-race *vs* other-race faces (Golby *et al.*, 2001; Lieberman *et al.*, 2005; Kim *et al.*, 2006) was absent when orthogonal group distinctions were made salient. This raises an important question that can be tested with MVPA: when race is irrelevant to group membership distinctions, does the face-sensitive FG still represent race, even though mean BOLD activity is driven by group membership, or is race 'erased' (i.e. no longer perceptually represented in the FG)?

In this research, we tested these possibilities by using MVPA to re-analyze an fMRI dataset collected by Van Bavel *et al.* (2011). Given that previous research has implicated the FG in the structural encoding of facial stimuli (Kanwisher *et al.*, 1997) and race perception (Golby *et al.*, 2001), we focused our analyses on the FG. Although our previous research suggests that race can be made irrelevant in certain contexts, we also have noted that 'race, like any physical or psychological property, may be represented in the brain, even when it is not exerting an influence on a specific mental process or task' (Van Bavel and Cunningham, 2011, p. 271). Therefore, we reasoned that MVPA would reveal representations of race in the FG even when univariate analyses find no differential effect of target race (Black *vs* White) on neural activity within the FG. We also analyzed a region of early visual cortex to determine whether representations of race in the FG reflected low-level perceptual information (e.g. color, contrast) fed forward from early visual cortex. Moreover, to determine that our findings were not simply a result of the entire brain responding to race, we also analyzed a control region outside the visual processing stream.

[1]To illustrate how this could occur, an analogy to the American presidential election process is useful. The boundaries of the United States are used to define the ROI, the individual states are the voxels, and the two candidates are the separate conditions. A univariate fMRI analysis is like the popular vote. In determining the results, the vote counters collapse across the tallies from the individual states and whichever candidate has the most overall votes is the winner. Multivariate techniques are similar to interpreting the meaning of the vote based on the pattern of states or districts that voted 'blue' or 'red'. The overall popular vote could be a statistical tie, similar to when there is no mean-level difference in neural activation; however, the pattern of 'blue' and 'red' states still provides interesting information about how the country voted (e.g. revealing strong regional preferences for different parties).

METHOD

Participants

As was the case for Van Bavel *et al.* (2011), data from 17 White participants (mean age = 20) were analyzed for this study. Each participant was paid $40 for completing the study and provided written informed consent prior to the start of the protocol. The session took place at the neuroimaging facility at Queen's University.

Procedure

Group assignment and learning

After consent was obtained, participants were led to a behavioral testing room. They were told that they would be assigned to a team: the Leopards or the Tigers, and that before beginning the scanning session, it was important for them to memorize the faces of the people who belonged to the two teams. Participants were randomly assigned to one of the teams and then completed two learning tasks to familiarize themselves with the members of each team (Van Bavel *et al.*, 2008; Van Bavel and Cunningham, 2009). Because the current dataset was described in a previously published study (Van Bavel *et al.*, 2011), we only report methodological details relevant to the present re-analysis.

During the first learning task, participants spent 3 min memorizing 16 male faces that were divided into two teams of eight (Leopards and Tigers). All the faces were presented simultaneously on the screen. Face stimuli were color images created in Photoshop and presented as 2×2.5 in at 72 pixels per inch. All faces had a neutral expression and were oriented according to the same forward-facing angle. Each team had an even number of Black and White faces, and assignment was fully counterbalanced so that no perceptual cues allowed participants to visually sort the faces into teams.

The second learning task was designed to reinforce their team affiliation and further strengthen their memory for the members of each team. It lasted ~13 min and was separated into two blocks. During each block, participants were asked to categorize faces one-at-a-time as members of the Leopard or Tiger team. To ensure that participants identified with their team, the participants also categorized a digital photograph of their own face. Participants did not view their own face in any of the subsequent parts of the study. During the first block of the second learning task, a label was used to remind participants whether each face was a Leopard or Tiger. Participants categorized eight in-group and eight out-group faces one time each and their own face three times, for a total of 19 trials. The team labels were removed during the second block, which forced participants to rely on their memory to accurately categorize the faces. After each trial in the second block, feedback indicated whether the response was correct and listed the correct team affiliation for each face. Participants categorized each in-group and out-group face three times and their own face three times during the second block, for a total of 51 trials.

Face categorization

After the learning and group assignment, participants were escorted to a Siemens 3T Tim Trio scanner, where they were positioned for the scanning session. All stimuli presented during the fMRI session were back-projected from an LCD projector to a clear screen at the back of the scanner bore. Participants were able to see these stimuli using a mirror mounted on top of the head coil (the visual angle of the stimuli was $\sim 8^\circ \times 6^\circ$). Stimuli were presented one-at-a-time in the center of an otherwise black screen. Participants completed a face categorization task that followed a mixed block/event-related design of five runs.

Each run comprised four randomly ordered blocks: two in-group categorization blocks and two out-group categorization blocks (Figure 1). During in-group categorization blocks, participants pressed

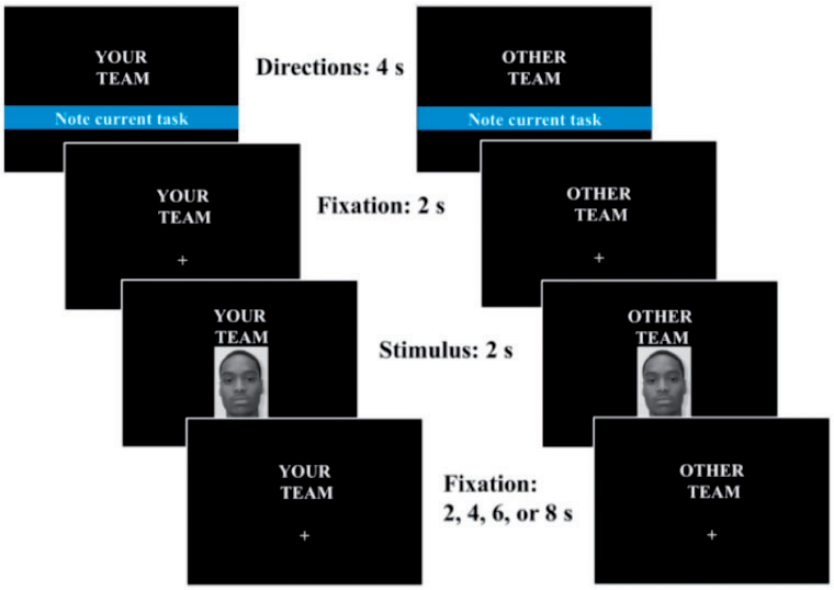

Fig. 1 Sample trials in the in-group categorization block (left) and out-group categorization block (right) during fMRI. Each block started with a directions screen (the top screen in the figure). After the directions screen, participants completed 12 trials. On each trial, participants hit a button if a randomly presented face (the third screen in the figure) was an in-group member (the left screens in the figure) or out-group member (the right screens in the figure) and then saw a fixation cross (the bottom screen in the figure). Each face appeared for 2 s, during which time participants responded with a button box in their right hand. To allow for estimation of the hemodynamic signal, fixation crosses appeared between faces for 2, 4 or 6 s (in pseudo-random order). After the completion of each block, directions for the next block appeared. Each of five runs contained two in-group categorization blocks and two out-group categorization blocks (counterbalanced).

a button only if the face was an in-group member. During out-group categorization blocks, participants pressed a button only if the face was an out-group member. Every block consisted of 12 trials, for a total of 240 trials. The block type was indicated for 4 s before each block began. The trials in each block were separated by a fixation cross that appeared for 2, 4 or 6 s (in a pseudo-random order). This jittered presentation allowed for modeling of the hemodynamic signal. Following the fixation cross, a face appeared for 2 s. The face was drawn from a pool of 24 faces. The pool contained eight in-group faces, eight out-group faces and eight novel faces of individuals who were unaffiliated with the in-group or out-group. Each face was presented twice in each run: once during in-group categorization and once during out-group categorization. Participants saw the unaffiliated faces for the first time during fMRI scanning. Faces were racially diverse such that half of the faces were White and half were Black (i.e. race was orthogonal to group membership).

Neuroimaging parameters, acquisition and preprocessing

Changes in the fMRI BOLD signal were measured using a single-shot gradient echo-planar pulse sequence [32 axial slices; 3.5 mm thick; 0.5 mm skip; echo time = 25 ms; repetition time (TR) = 2000 ms; in-plane resolution = 3.5 × 3.5 mm; matrix size = 64 × 64; field of view = 224 mm]. Preprocessing was done with SPM8 (Wellcome Department of Cognitive Neurology, London, UK). Data were realigned to the first image and corrected for slow signal drift with a 128 s high-pass filter. The time series from each voxel was de-trended to remove linear and quadratic trends, and *z*-scored to normalize the time series to have a mean of zero and a variance of one. Condition onsets were adjusted for the lag in hemodynamic response function by shifting all block-onset timings by three volumes (6 s).

Localization of the fusiform gyri and control regions of interest

To localize the ROIs, we first performed a within-participant analysis with a voxel-wise general linear model. The model comprised fourteen boxcar waveforms representing the experimental conditions: for two different tasks, six regressors modeling Black and White faces that were part of the in-group, out-group or unaffiliated with either group, plus two regressors to model direction screens and the duration of the rest period (comprising only a fixation cross). We then computed the contrast of all faces *vs* rest.[2] For each participant, this contrast contained a balanced number of blocks with the same number of Black and White faces.

On the basis of this contrast, we located a face sensitive region of the FG bilaterally. The peak of the activation in the FG defined the center of two 10 mm diameter sphere-shaped ROIs (one per hemisphere). We also created two other ROIs. One ROI comprised an area of each participant's early visual cortex (VC) that approximated primary visual cortex. We included an ROI for VC because this brain region is sensitive to low-level visual differences, including color perception (Brouwer and Heeger, 2009) and increased attention (Kastner *et al.*, 1999), but not the higher-order social significance of race. The additional ROI was a size-matched control region (CTR) in an area of the medial orbitofrontal cortex that was not face-sensitive according to our face localizer (see also Kaul *et al.*, 2011). The VC and CTR ROIs each comprised one medially located sphere with equal volume (12.6 mm diameter). The central coordinates of the ROIs for each participant are listed in Table 1.

Univariate analysis

Previously published results obtained from these data using a univariate analysis (Van Bavel *et al.*, 2011) found that the mean BOLD signal in the FG did not significantly differ between Black and White faces when novel group membership was the relevant categorical dimension. The goal of the current univariate analysis was to replicate this previous finding using the same preprocessing steps and

[2]Although a functional Fusiform Face Area (FFA) localizer was collected (Van Bavel *et al.*, 2011), for many participants this localizer did not yield enough voxels to conduct MVPA. MVPA requires multiple voxels for each ROI for each participant. We therefore used the alternate functional localizer described in the text. Univariate analyses replicated across both localizers.

Table 1 Central coordinates of the ROIs for each participant

Participant	FG (L)	FG (R)	VC	CTR
1	[−30, −13, −12]	[37, −27, −4]	[1, −50, 0]	[4, 56, 23]
2	[−32, −2, −2]	[32, −6, −6]	[−2, −52, −3]	[3, 62, 22]
3	[−38, −27, −13]	[28, −13, −16]	[−2, −41, 21]	[6, 43, 24]
4	[−23, −29, −6]	[22, −25, −6]	[−2, −50, −28]	[5, 59, 28]
5	[−36, −29, −22]	[48, −22, −26]	[−9, −32, 8]	[6, 53, 20]
6	[−36, −29, −22]	[36, −2, −22]	[−2, −55, −19]	[6, 48, 26]
7	[−33, −15, −17]	[40, −30, −10]	[5, −55, −1]	[4, 31, 29]
8	[−31, −27, −9]	[42, −27, −6]	[14, −59, −2]	[8, 48, 16]
9	[−39, −29, −4]	[32, −19, −7]	[0, −53, 8]	[3, 33, 26]
10	[−35, −37, −10]	[33, −30, −13]	[0, −54, −10]	[7, 48, 8]
11	[−43, −6, −21]	[39, −13, −25]	[3, −46, 7]	[6, 35, 13]
12	[−32, −14, −28]	[34, −14, −35]	[14, −57, −25]	[3, 21, −2]
13	[−35, −10, −2]	[39, −20, −5]	[−3, −56, −6]	[9, 43, 23]
14	[−30, −12, −27]	[33, −18, −28]	[−2, −57, −28]	[4, 48, 4]
15	[−33, −6, −16]	[40, −1, −15]	[−9, −29, −11]	[4, 58, 13]
16	[−38, −16, −19]	[40, −13, −18]	[−2, −56, −8]	[3, 52, 17]
17	[−38, −17, −26]	[45, −13, −18]	[−10, −60, −5]	[11, 45, 11]

Note: FG (L), left fusiform gyrus; FG (R), right fusiform gyrus; VC, early visual cortex; CTR, control region in the prefrontal cortex.

voxels as the MVPA. The only exception was that data were spatially smoothed prior to the univariate analysis, whereas unsmoothed data were used in the MVPA.[3] After smoothing the data, the BOLD responses to Black and White faces (irrespective of group membership) were calculated by averaging the signal from voxels within each ROI (FG, VC and CTR). We then compared these mean BOLD values collapsed across all subjects.

Multivariate analysis

The preprocessed data without spatial smoothing from the five experimental runs were analyzed using the MATLAB routines provided in the Princeton MVPA Toolbox (www.csbmb.princeton.edu/mvpa). To determine classification accuracies, only classification with unseen and independent test data was considered, using a leave-one-session-out cross-validation method (Mur *et al.*, 2009; Pereira *et al.*, 2009). In the actual classification step, we used a Gaussian Naïve Bayes classifier algorithm (see Mitchell *et al.*, 2004) within the MVPA toolbox.

Classification accuracies were averaged across the five cross-validations for each ROI for each participant. Thus, for each participant, this procedure yielded exactly one mean classification accuracy per ROI (i.e. 17 total observations per ROI). We then used paired *t*-tests to assess significant differences in decoding accuracies from chance (two categories = 50% chance) and a control-baseline defined by the classification accuracy within CTR. We also examined whether our results were robust across hemisphere, block type (in-group or out-group categorization) and group (in-group, out-group, unaffiliated).

In order to evaluate the probability that the classification was driven by over-fitting of arbitrary patterns of spatial correlations in the data, we used the shuffle control routine in the MVPA toolbox to conduct a permutation test that involved reshuffling training labels for each round of the cross-validation (Kaul *et al.*, 2011; Mur *et al.*, 2009). If the null assumption that classification is driven by chance were true, similar results should be obtained if labels indicating race during training were shuffled randomly. We expected the resulting distribution of classification accuracies to confirm the expected distribution for chance prediction (two categories = 50% chance).

[3]The data for the univariate analyses were spatially smoothed to maximize the signal-to-noise ratio (Mikl *et al.*, 2008). Due to the possibility that spatial smoothing can remove fine-grained pattern information, we did not spatially smooth the data prior to the MVPA (Kriegeskorte *et al.*, 2006; Mur *et al.*, 2009, but see Kamitani and Sawahata, 2010; Op de Beeck, 2010).

RESULTS

Behavioral results

To assess behavioral responses during fMRI, we used paired *t*-tests to compare participants' reaction time (ms) and accuracy to in-group *vs* out-group blocks of the Face Categorization Task. Both blocks were relatively difficult (mean accuracy = 58.0%, where chance = 33.3%). However, participants were faster, $t(16) = 2.90$, $P < 0.01$ and more accurate, $t(16) = 3.07$, $P < 0.01$, to categorize faces during the in-group (1223 ms; 62.0%) *vs* the out-group (1306 ms; 54.1%) blocks.

Univariate results

Replicating previously published analyses on the current dataset (Van Bavel *et al.*, 2011), paired *t*-tests indicated that the mean BOLD signal in all three ROIs did not significantly differ between Black and White faces (P's > 0.47). Also replicating past findings, when collapsing across race, the mean BOLD signal was significantly greater to in-group *vs* out-group faces in the FG, $t(16) = 2.4$, $P < 0.05$ and VC, $t(16) = 2.4$, $P = 0.05$, but not the control region, $t(16) = 1.4$, n.s.

MVPA results

In line with the view that race is represented in the FG even when it is not associated with racial bias in mean BOLD signal and is not explicitly relevant to categorization (see Van Bavel and Cunningham, 2011), MVPA indicated that race could be decoded better than chance in the FG, 56.6%, $t(16) = 6.11$, $P < 0.01$ and in the VC, 52.3%, $t(16) = 2.47$, $P < 0.05$. Importantly, race was not represented in the control area, 49.8%, $t(16) = -0.30$, n.s., suggesting that these effects were not due to global patterns in race decoding. Figure 2A shows the mean decoding accuracies, averaged across participants. We then defined the distribution of decoding results from the control region (CTR) as an alternate baseline (instead of 50% chance). Testing against this alternate baseline, we replicated race decoding in FG, $t(16) = 5.20$, $P < 0.01$ and in VC, $t(16) = 2.10$, $P = 0.05$.

The FG data reported above were collapsed across hemispheres. However, it is well documented that the FG shows a degree of asymmetry in its response to faces (Kanwisher *et al.*, 1997). To evaluate any possible differences in hemispheric classification accuracy, we repeated the analysis in the FG for each hemisphere. Right and left FG successfully predicted facial race at similar levels to that seen when analyzed together, right FG: 55.3%, $t(16) = 5.10$, $P < 0.01$, left FG: 55%, $t(16) = 6.12$, $P < 0.01$. There were no significant differences when comparing classification accuracies of left and right FG across all participants $t(16) = -0.23$, n.s.

Next, we examined the possibility that prediction accuracy in FG might reflect low-level visual information propagated from early visual cortices. To this end, we compared the decoding results of FG and VC. Facial race decoding was significantly higher in FG than in VC, $t(16) = 2.98$, $P < 0.05$, suggesting relatively greater race-relevant information in the neural pattern in FG relative to VC. Moreover, decoding accuracies from VC and FG were not significantly correlated, $r = 0.17$, $P = 0.52$, further corroborating the relative independence of information represented in the two regions. This finding suggests that when categorizing faces on the basis of group membership, race processing not only involves low-level visual features (e.g. skin color), but also additional information (e.g. configural properties).

During data collection, participants performed one of two group membership tasks, reporting whether the faces belonged to the

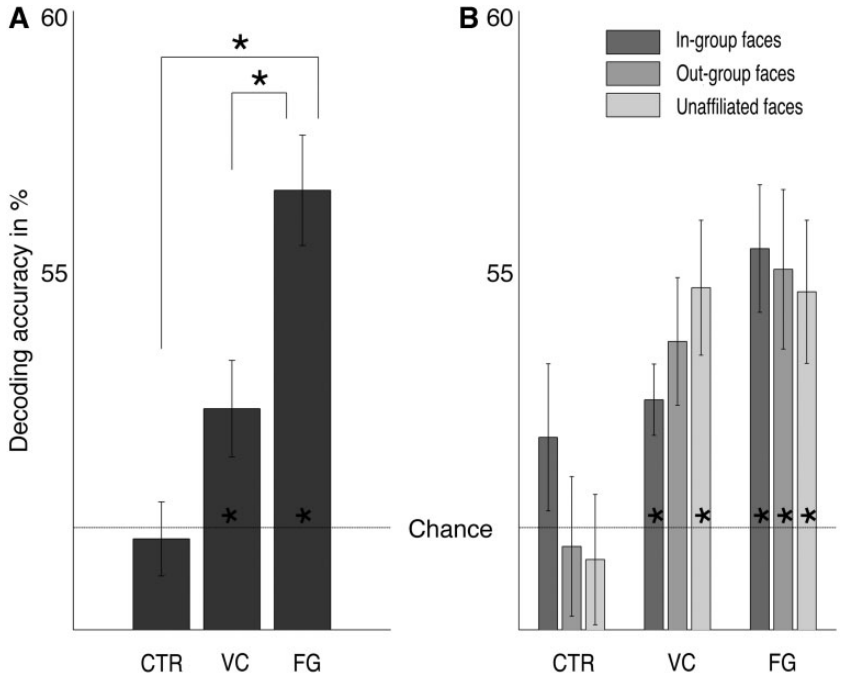

Fig. 2 (A) Mean race decoding accuracy in the FG, VC, and CTR. (B) Mean race decoding accuracy for each of the three subgroups: in-group, out-group and unaffiliated faces. *$P < .05$.

in-group (yes/no) or out-group (yes/no). Although participants generally responded faster to the faces when performing the in-group task, there was no theoretical reason that race should be decoded differently during these two tasks. Thus, to test this reasoning, we repeated our MVPA analysis separately for each of the two tasks. As predicted, for both tasks the results replicated the combined analysis, and there were no significant differences between the tasks, FG: 58%/56.1%, $t(32) = 1.9$, n.s.; CTR: 51.2%/50.7%, $t(32) = -1.71$, n.s.

To ensure that race decoding was not dependent on group membership, we repeated the analysis within each stimulus subgroup that the study design offered (in-group, out-group and unaffiliated faces). As depicted in Figure 2B, the pattern of the FG result was similar in all three groups, in-group: 55.5%, $t(16) = 4.38$, $P < 0.01$; out-group: 55.1%, $t(16) = 3.24$, $P < 0.05$; unaffiliated: 54.6%, $t(16) = 3.29$, $P < 0.05$. Results in the control region were again at chance prediction, in-group: 51.8%, $t(16) = 1.23$, n.s.; out-group: 49.6%, $t(16) = -0.27$, n.s.; unaffiliated: 49.4%, $t(16) = -0.49$, n.s. Three separate analyses of variance that tested for differences between the three face-categories in each ROI did not reveal any significant results, FG: $F(48) = 0.09$, n.s.; VC: $F(48) = 0.95$, n.s.; CTR: $F(48) = 0.93$, n.s.

Finally, to rule out the possibility that successful race decoding was driven by stimulus-independent spatial correlations in the data (independent of the race of a face) and over-fitting arbitrary patterns of spatial correlations in the data, we carried out a shuffle-control test (Mur *et al.*, 2009; Kaul *et al.*, 2011). If race decoding were driven by chance, similar results should be obtained if labels indicating the conditions during training were shuffled randomly. To test this possibility, we ran a separate analysis using the shuffle-control routine with the MVPA toolbox, in which labels during training were re-shuffled for each round of the cross-validation. We expected the resulting distribution of decoding accuracies to confirm the expected distribution for chance prediction (two categories = 50% chance). The result confirmed the distribution of decoding accuracies expected under the null hypothesis, VC: 50.5%, $t(16) = 0.82$, n.s.; FG, 50.5%, $t(16) = 0.63$, n.s.; CTR, 50.1%, $t(16) = 0.09$, n.s.

DISCUSSION

In this research, we examined the underlying neural representations of race in the FG and early visual cortex using MVPA, an analytic technique that can identify category-based neural representations in the absence of mean-level differential activity between categories (see Kaul *et al.*, 2011). As predicted, multivariate analyses of patterns of neural activity within the FG could decode the race of faces above chance even when univariate analyses were not able to detect mean-level race differences in the FG. Importantly, race decoding in a size-matched control region was not significantly different from chance, suggesting that decoding accuracy for facial race was not due to potential confounds, such as a general increase in blood flow. Moreover, race was represented in the FG to a greater extent than early visual cortex, which suggests that the FG effect did not merely reflect low-level perceptual information (e.g. color, contrast) propagated from early visual cortex.[4] The results of this research indicate that *patterns* of activation within the FG continue to encode race even when mean FG activation is driven by other factors.

We speculate that whereas early visual cortex is largely sensitive to race due to low-level visual cues (e.g. skin color), the FG, as a brain area implicated in higher-order visual processing, represents race in our study because race provides an individuating cue that facilitates categorization on the task-relevant group membership dimension. In the task in our study, group membership of each face was not indicated by a perceptual cue and instead had to be encoded in memory. Thus, to successfully complete the task, participants needed to retrieve each target's group membership from memory. To the extent that a target's race helped participants access information about the target in memory, race representation may have facilitated the group membership categorization. We mention this as a potential explanation for race representation in the FG in our study; however,

[4]Although greater classification accuracies of the FG *vs* early visual cortex suggest that the reported FG effects do not simply reflect low-level visual processing, it is important to note that we are not able to rule out the possibilities that these areas differentially represent information according to an unknown non-linear structure or that our results are dependent on the particular resolution of our fMRI data.

further research will be necessary to examine this possibility (see Kaul *et al.*, 2012).

It is noteworthy that although our MVPA analyses suggest that race is represented in the FG, behavioral research using the same novel group, mixed-race paradigm has demonstrated that evaluations and memory for faces are characterized by biases in group membership, not race (Van Bavel and Cunningham, 2009; Van Bavel *et al.*, 2012). Indeed, this behavioral pattern matches the univariate results. Race is represented in the FG, but the mean response of the FG does not reflect racial differences among the target faces. Perhaps, as we posit above, race facilitates activation of relevant non-perceptual group information from memory, but once the group information is activated, the race information is no longer useful for task completion. Thus, the possibility arises that race is perceptually represented in the brain, even when it is functionally erased in terms of biased evaluations and behavior (Cunningham *et al.*, 2007; Van Bavel and Cunningham, 2009; Van Bavel *et al.*, 2012a,b).

Many social psychological perspectives suggest that group differences can be bridged by minimizing group distinctions and appealing to higher-order common identities (Allport, 1954; Sherif *et al.*, 1961; Gaertner *et al.*, 1993). The appeal of this research has contributed to the emergence of 'colorblind' initiatives, which assume that acknowledging race is harmful to harmonious intergroup relations. Our findings suggest, however, that the brain may detect and represent race in contexts where behaviors are not negatively impacted by racial representations. Thus, it appears that the way the brain actually processes race is consistent with policies that recognize both that phenotypic differences between races are difficult to ignore, and that noticing racial differences does not necessarily mean that people will be evaluated or treated poorly. In fact, policies that embrace recognition of racial diversity have been shown to outperform policies that encourage people to ignore racial differences (Apfelbaum *et al.*, 2010).

Returning to Martin Luther King Jr, although the words 'perceived' and 'judged' are often used interchangeably, it is notable that he dreamt that his children would not be 'judged' by the color of their skin. Perhaps King recognized that 'seeing' race is not inherently problematic for race relations. It is what the mind subsequently does with this information that matters.

REFERENCES

Allport, G.W. (1954). *The Nature of Prejudice*. Cambridge, MA: Perseus Books.

Apfelbaum, E.P., Pauker, K., Sommers, S.R., Ambady, N. (2010). In blind pursuit of racial equality? *Psychological Science*, *21*, 1587–92.

Brouwer, G.J., Heeger, D.J. (2009). Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience*, *29*, 13992–4003.

Caldara, R., Thut, G., Servoir, P., Michel, C.M., Bovet, P., Renault, B. (2003). Face versus non-face object perception and the 'other-race' effect: a spatio-temporal event-related potential study. *Clinical Neurophysiology*, *114*, 515–28.

Chiu, Y.C., Esterman, M., Rosen, H., Yantis, S. (2011). Decoding task-based attentional modulation during face categorization. *Journal of Cognitive Neuroscience*, *23*, 1198–204.

Cosmides, L., Tooby, J., Kurzban, R. (2003). Perceptions of race. *Trends in Cognitive Sciences*, *7*, 173–9.

Cunningham, W.A., Zelazo, P.D., Packer, D.J., Van Bavel, J.J. (2007). The Iterative Reprocessing Model: a multi-level framework for attitudes and evaluation. *Social Cognition*, *25*, 736–60.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.J. (1995). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, *2*, 189–210.

Gaertner, S.L., Dovidio, J.F., Anastasio, P.A., Bachman, B.A., Rust, M.C. (1993). The common ingroup identity model: recategorization and the reduction of intergroup bias. *European Review of Social Psychology*, *4*, 1–26.

Golby, A.J., Gabrieli, J.D.E., Chiao, J.Y., Eberhardt, J.L. (2001). Differential fusiform responses to same- and other-race faces. *Nature Neuroscience*, *4*, 845–50.

Haynes, J.D., Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*, 523–34.

Hehman, E., Mania, E.W., Gaertner, S.L. (2010). Where the division lies: common ingroup identity moderates the cross-race effect. *Journal of Experimental Social Psychology*, *46*, 445–8.

Ishai, A., Ungerleider, L.G., Martin, A., Schouten, J.L., Haxby, J.V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, *96*, 9379–84.

Ito, T.A., Urland, G.R. (2003). Race and gender on the brain: electrocortical measures of attention to the race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology*, *85*, 616–26.

Kamitani, Y., Sawahata, Y. (2010). Spatial smoothing hurts localization but not information: pitfalls for brain mappers. *Neuroimage*, *49*, 1949–52.

Kanwisher, N., McDermott, J., Chun, M. (1997). The Fusiform Face Area: a module in human extrastriate cortex specialized for the perception of faces. *Journal of Neuroscience*, *17*, 4302–11.

Kastner, S., Pinsk, M.A., De Weerd, P., Desimone, R., Ungerleider, L.G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, *22*, 751–61.

Kaul, C., Rees, G., Ishai, A. (2011). The gender of face stimuli is represented in multiple regions in the human brain. *Frontiers in Human Neuroscience*, *4*, 238.

Kim, J.S., Yoon, H.W., Kim, B.S., Jeun, S.S., Jung, S.L., Choe, B.Y. (2006). Racial distinction of the unknown facial identity recognition mechanism by event-related fMRI. *Neuroscience Letters*, *3*, 279–84.

Kriegeskorte, N., Goebel, R., Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*, 3863–8.

Kurzban, R., Tooby, J., Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences*, *98*, 15387–92.

Lieberman, M.D., Hariri, A., Jarcho, J.M., Eisenberger, N.I., Bookheimer, S.Y. (2005). An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. *Nature Neuroscience*, *8*, 720–2.

Macrae, C.N., Quadflieg, S. (2010). Perceiving people. In: Gilbert, D.T., Fiske, S.T., Lindzey, G., editors. *The Handbook of Social Psychology*. New York, NY: McGraw-Hill.

Mikl, M., Marecek, R., Hlustík, P., et al. (2008). Effects of spatial smoothing on fMRI group inferences. *Magnetic Resonance Imaging*, *26*, 490–503.

Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X. (2004). Learning to decode cognitive states from brain images. *Machine Learning*, *57*, 145–75.

Mur, M., Bandettini, P.A., Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience*, *4*, 101–9.

Natu, V., Raboy, D., O'Toole, A.J. (2010). Neural correlates of own- and other-race face perception: spatial and temporal response differences. *Neuroimage*, *54*, 2547–55.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*, 424–30.

Op de Beeck, H. (2010). Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? *Neuroimage*, *49*, 1943–8.

Pereira, F., Mitchell, T., Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, *45*, S199–209.

Puce, A., Allison, T., Gore, J.C., McCarthy, G. (1995). Face-sensitive regions in human extrastriate cortex studied by functional MRI. *Journal of Neurophysiology*, *74*, 1192–1199.

Sergent, J., Ohta, S., MacDonald, B. (1992). Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain*, *115*, 15–36.

Sherif, M., Harvey, O.J., White, B.J., Hood, W.R., Sherif, C.W. (1961). *Intergroup Conflict and Cooperation: The Robbers Cave Experiment*. Norman, OK: University of Oklahoma Book Exchange.

Sidanius, J., Pratto, F. (1999). *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression*. New York, NY: Cambridge University Press.

Van Bavel, J.J., Cunningham, W.A. (2009). Self-categorization with a novel mixed-race group moderates automatic social and racial biases. *Personality and Social Psychology Bulletin*, *35*, 321–35.

Van Bavel, J.J., Cunningham, W.A. (2011). A social neuroscience approach to self and social categorisation: a new look at an old issue. *European Review of Social Psychology*, *21*, 237–84.

Van Bavel, J.J., Packer, D.J., Cunningham, W.A. (2008). The neural substrates of in-group bias: a functional magnetic resonance imaging investigation. *Psychological Science*, *19*, 1131–9.

Van Bavel, J.J., Packer, D.J., Cunningham, W.A. (2011). Modulation of the Fusiform Face Area following minimal exposure to motivationally relevant faces: evidence of in-group enhancement (not out-group disregard). *Journal of Cognitive Neuroscience*, *23*, 3343–54.

Van Bavel, J.J., Swencionis, J., O'Connor, R., Cunningham, W.A. (2012a). Motivated social memory: belonging needs moderate the own-group bias in face recognition. *Journal of Experimental Social Psychology*, *48*, 707–13.

Van Bavel, J.J., Xiao, Y.J., Cunningham, W.A. (2012b). Evaluation as a dynamic process: moving beyond dual system models. *Social and Personality Psychology Compass*, *6*, 438–54.