A Cognitive Neuroscience of Social Groups

A dissertation presented

by

Juan Manuel Contreras

to

the *Department of Psychology*

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Psychology

Harvard University

Cambridge, Massachusetts

May 2013

Dissertation Advisor: Professor Mahzarin Banaji Juan Manuel Contreras

Dissertation Advisor: Professor Jason Mitchell

A Cognitive Neuroscience of Social Groups

ABSTRACT

We used functional magnetic resonance imaging to investigate how the human brain processes information about social groups in three domains. *Study 1: Semantic knowledge.* Participants were scanned while they answered questions about their knowledge of both social categories and non-social categories like object groups and species of nonhuman animals. Brain regions previously identified in processing semantic information are more robustly engaged by nonsocial semantics than stereotypes. In contrast, stereotypes elicit greater activity in brain regions implicated in social cognition. These results suggest that stereotypes should be considered distinct from other forms of semantic knowledge. *Study 2: Theory of mind.* Participants were scanned while they answered questions about the mental states and physical attributes of individual people and groups. Regions previously associated with mentalizing about individuals were also robustly responsive to judgments of groups. However, multivariate searchlight analysis revealed that several of these regions showed distinct multivoxel patterns of response to groups and individual people. These findings suggest that perceivers mentalize about groups in a manner qualitatively similar to mentalizing about individual people, but that the brain nevertheless maintains important distinctions between the representations of such entities. *Study 3: Social categorization.* Participants were scanned while they categorized the sex and race of unfamiliar Black men, Black women, White men, and White women.

Multivariate pattern analysis revealed that multivoxel patterns in FFA—but not other face-selective brain regions, other category-selective brain regions, or early visual cortex—differentiated faces by sex and race. Specifically, patterns of voxel-based responses were more similar between individuals of the same sex than between men and women, and between individuals of the same race than between Black and White individuals. These results suggest that FFA represents the sex and race of faces. Together, these three studies contribute to a growing cognitive neuroscience of social groups.

TABLE OF CONTENTS

To Armando Cardozo, Ph.D. and Manuel E. Contreras, Ph.D.

ACKNOWLEDGMENTS

For advice and assistance, I thank Regan Bernhard, Susanna Carmona, Michael Cohen, Amrita Dani, Katharine Dobos, Dave Johnson, Eshin Jolly, Anna Leshinskaya, Grace Mahoney, Stephanie McMains, Ross Mair, Joe Moran, Tammy Moran, Brandi Newell, Jess Schirmer, Diana Tamir, Jennifer Tao, Emma Thompson, Mark Thornton, Ahn Ton, Adam Waytz, and Rachel Wong.

For financial support, I thank the Department of Education for a Jacob K. Javits Fellowship, the National Science Foundation for a Graduate Research Fellowship, and the Department of Psychology of Harvard University for a grant from the Stimson Fund. The studies in this dissertation were funded by Harvard University and a grant from the National Science Foundation (BCS 0642448) to Mahzarin Banaji and Jason Mitchell.

For advice on this dissertation, I thank Alfonso Caramazza, Mahzarin Banaji, Joshua Greene, and Jason Mitchell.

For scientific wisdom, thoughtful guidance, relentless help, tireless patience, and moral support, I thank Mahzarin Banaji and Jason Mitchell.

INTRODUCTION

Human beings are intensely social animals. Every one of us is member of a myriad of social groups that shape most, if not all, aspects of our lives. From how we surf the web (men and women differ in amount and kind of internet use; e.g., Joiner et al., 2012) to the political beliefs that we hold (parents transmit political preferences to their children via genes; e.g., Kandler, Bleidorn, & Riemann, 2012), social groups shape the way we act and think. But biological groups like sex and family are not the only groups to which we belong. We join groups to support the same sports team, root for the same political candidate, worship the same deity, and listen to the same type of music. The formation of social groups on non-biological bases is a universal of human nature (Wilson, 2012). As a result, we readily organize and thrive in groups that are orders of magnitude more complex than those of other social primates (Hill & Dunbar, 2003).

Given the importance of social groups to humans, social psychologists have investigated the relationship between an individual's psychology and group dynamics extensively—how we think about groups, how being in a group influences our behavior, how we interact with members of other groups, how intergroup conflict arises, etc. (for reviews, see Dovidio & Gaertner, 2010; Hackman & Katz, 2010; Yzerbyt & Demoulin, 2010). Recently, social psychologists extended their study of the social nature of humans to its neural basis (for reviews, see Decety & Cacioppo, 2011; Todorov, Fiske, & Prentice, 2011). Using neuroimaging techniques like functional magnetic resonance imaging (fMRI), social psychologists and, now, cognitive neuroscientists, study the role that different regions of the human brain play in sustaining the complex social cognition that enables us to be functional social beings.

Increasingly, these scientists have brought the methods of cognitive neuroscience to bear on questions about the psychology of how we think about other individuals as members of other groups (for reviews, see Cunningham & Van Bavel, 2009; Ito & Bartholow, 2009; Kubota, Banaji, & Phelps, 2012; Van Bavel & Cunningham, 2009). This research has started to shed light on how we perceive other people as members of social groups (e.g., as African Americans). In doing so, these studies have started to demonstrate that neuroimaging can furnish insights into intergroup cognition.

For example, many experiments have reported robust amygdala activity to the perception of the faces of racial outgroups (Amodio, Harmon-Jones, & Devine, 2003; Cunningham, Raye, & Johnson, 2004; Lieberman, Hariri, Jarcho, Eisenberger, & Bookheimer, 2005; Phelps et al., 2000; Richeson et al., 2003; Ronquillo et al., 2007; Wheeler & Fiske, 2005). This finding has been interpreted as increased vigilance of members of other races, presumably because people can dislike and, in turn, distrust members of other races. Indeed, the degree of amygdala activity that White participants show when they view Black faces is positively correlated with the amount of implicit anti-Black prejudice that they have (Beer et al., 2008; Cunningham, et al., 2004; Phelps, et al., 2000; Platek & Krill, 2009).

As another example, other experiments find robust ventral striatum activity to the perception of individuals who are high in social status, whether this status is obtained in a contrived economic game or from their actual socioeconomic background (Ly, Haynes, Barter, Weinberger, & Zink, 2011; Zink et al., 2008). Given the role of the ventral striatum in stimulus valuation (for review, see Ikemoto & Panksepp, 1999), this finding has been interpreted as a possible neural marker of the high value that we attribute to people with significant social status.

However, current studies on the cognitive neuroscience of social groups have three important limitations. First, they have not examined the functional neuroanatomy of perceiving and thinking about groups *qua* groups. That is, these studies have placed exclusive focus on how we think about individuals as members of other groups, but they have not investigated how we think about the groups themselves. So, for example, though we know that the perception of faces from individuals of other races influences amygdala response, we do not know how the retrieval of knowledge about African Americans as a whole differs from the retrieval of similar knowledge about nonsocial categories like object groups.

Second, current research has not addressed many important domains of intergroup cognition. For example, we often attribute mental states to groups of people (Jones, 2010) despite the fact that, though members of groups may have minds, groups themselves do not. Nonetheless, these attributions are real and they influence how we evaluate group groups as well as their members (e.g., Waytz & Young, 2012). Though cognitive neuroscientists have identified a set of brain regions that are reliably engaged during inferences about mental states (for reviews, see Frith & Frith, 2006; Mitchell, 2009a, 2009b; Saxe, 2006), we do not know if these brain regions also play a role in attributions of mind to groups and, if so, whether they differentiate between inferences about the mental states of groups and individuals.

Finally, current studies have focused almost exclusively on identifying which brain regions are involved in different aspects of intergroup cognition, but they have not attempted to identify brain regions that contain distinct representations of groups and individuals or distinct representations of different social groups. For example, though previous research has shown that fusiform face area (FFA), a face-selective portion of the fusiform gyrus (for review, see Kanwisher & Yovel, 2006), has a stronger response to same- than other-race faces (Golby, Gabrieli, Chiao, & Eberhardt, 2001; Lieberman, et al., 2005), it is not yet clear whether this brain region represents the race of faces.

The studies that comprise this dissertation start addressing the three limitations of current research on the cognitive neuroscience of social groups. All studies explore three domains of intergroup cognition that have not yet received significant attention by social neuroscientists: Semantic knowledge, theory of mind, and social categorization. Additionally, studies 1 and 2 investigate the functional neuroanatomy of thinking about groups rather than thinking about individuals as group members. Finally, studies 2 and 3 include experiments that use multivoxel pattern analysis (MVPA) to begin to understand the neural representations that distinguish individuals from groups and different types of social groups from each other.

Study 1 presents an fMRI experiment that suggests that semantic knowledge about groups of people, or stereotypes, and semantic knowledge about object categories have dissociable neural correlates. Brain regions previously identified in processing semantic information are more robustly engaged by nonsocial semantics than stereotypes. In contrast, stereotypes elicit greater activity in brain regions implicated in social cognition. Study 2 presents two fMRI experiments that suggest that theory of mind about individuals and theory of mind about groups have a similar neural basis. Also, this study finds evidence that these brain regions have distinct representations of groups and individuals during inferences about mental states. Finally, Study 3 presents an fMRI experiment that suggests that FFA represents the sex and race of faces, differentiating faces by sex and race at the level of multivoxel patterns.

Together, these three studies contribute to a growing cognitive neuroscience of social groups. This dissertation closes with a discussion of future research on how the brain processes information about social groups.

STUDY 1

DISSOCIABLE NEURAL CORRELATES OF STEREOTYPES AND OTHER FORMS OF SEMANTIC KNOWLEDGE

Contreras, J. M., Banaji, M. R., & Mitchell, J. P. (2012). Dissociable neural correlates of stereotypes and other forms of semantic knowledge. *Social Cognitive and Affective Neuroscience*, *7*(7), 764-770.

INTRODUCTION

Over the course of a day, we encounter a large number of objects. Even before we leave our homes in the morning, we have already interacted with dozens of objects, from alarm clocks to armoires, beds to belts, and cups to chairs. These interactions require the capacity to distinguish these objects from each other by recognizing their unique physical features, understanding the discrete functions they serve, and recalling the correct procedure for their use (how to set the alarm clock to snooze, the order in which to put on one's socks and shoes, etc.).

Philosophers and psychologists have long posited that such knowledge is necessarily organized around categories that distinguish among different sets of entities (Aristotle, 1975; Kant, 1781/2003; Medin & Smith, 1984; Murphy, 2002; Smith & Medin, 1981). Categories obviate the need to repeatedly work out what to expect from each object, by allowing perceivers to instead make use of generalized knowledge about a whole class of entities. For example, by recognizing a particular object as an instance of the category "microwave ovens," one gains immediate access to a wealth of additional information about it—such as that it can be used to heat food, cannot accommodate metal pots, and will probably have a button marked "defrost"—without the need to discover each of these features anew.

Psychologists have held that categories not only organize our understanding of inanimate objects, but likewise guide interactions with the myriad individuals with whom we come into daily contact (Allport, 1954). We readily categorize other individuals into a wide range of social groups, such as those based on gender, race and ethnicity, age, occupation, place of origin, socioeconomic class, and so on. Much as recognizing a particular object as a member of a general category provides useful information about that object "for free," categorizing a particular individual as a member of a social category (e.g., "men"; "New Yorkers") gives us ready access to the likely characteristics of this person (e.g., “he will probably be an aggressive driver”). Typically, the information that derives from social categorization is referred to as a *stereotype*—the inferences and assumptions made about a particular person as a consequence of categorizing him into one or another social group.

Given that stereotypes serve much the same function as other forms of category-based knowledge, many researchers have naturally assumed that stereotypes merely reflect ordinary semantic knowledge about a particular class of entities—other people. For example, stereotypes have been described as the “perception of social objects (e.g., groups) that is in principle little different to categorization and perception of other ‘physical’ objects” (Spears, Oakes, Ellemers, & Haslam, 1997, p. 3); “not essentially different from other cognitive structures or processes” (Hamilton, 1981, p. 28); and “rooted in the ordinary mechanisms of perception and categorization” (Banaji & Bhaskar, 1999, p. 144). At the same time, a number of researchers have argued that stereotypes may instead be a unique form of semantic knowledge (e.g., Ostrom, 1984). Social groups are generally more complex than categories of nonsocial objects (Cantor & Mischel, 1979; Wattenmaker, 1995), individuals typically belong to several social categories simultaneously (Lingle, Altom, & Medin, 1984; Schneider, 2004), and stereotypes often evoke more emotion than other forms of semantic knowledge (Norris, Chen, Zhu, Small, & Cacioppo, 2004). Given these distinct aspects of social knowledge, some researchers have suggested that social knowledge may require specialized forms of cognitive processes that distinguish it from other forms of semantics.

Are stereotypes a typical form of semantic knowledge or do they represent a unique form of knowledge about the (social) world? Historically, it has been difficult to adjudicate between these competing accounts of stereotyping. However, recent findings regarding the neural basis of semantics offer a novel strategy for addressing this question. Over the past decade, researchers have consistently demonstrated that a small number of left-lateralized brain regions subserve the retrieval, selection, and integration of information from semantic memory; specifically, inferior frontal gyrus and inferotemporal cortex (for reviews, see Bookheimer, 2002; Joseph, 2001; Martin, 2001). For example, participants show greater hemodynamic activity in left inferior frontal gyrus when thinking about the meaning of a word than when they consider its perceptual characteristics (such as whether the word is written in uppercase letters; e.g., Poldrack et al., 1999), and regions of left inferotemporal cortex have routinely been observed when participants name or simply view categories of objects (Martin & Chao, 2001). Moreover, people with damage to these regions often demonstrate selective impairments in semantic memory, such as an inability to name common objects or define familiar words (Baldo & Shimamura, 1998; Caramazza & Shelton, 1998; Hodges, Patterson, Oxbury, & Funnell, 1992). Moreover, people with damage to anterior temporal lobes can show similar semantic deficits (Patterson, Nestor, & Rogers, 2007).

To the extent that stereotypes are part of general semantics, these same regions should contribute to the retrieval of knowledge about the attributes of social groups. That is, if knowledge about social groups (i.e., stereotypes) does not differ significantly from knowledge about groups of objects, inferior frontal gyrus and inferotemporal cortex should be engaged when perceivers consider the typical features of social categories, such as those based on race, national origin, or occupation. From the point of view of the neural processes involved, thinking about Dutch ovens, Swedish meatballs, or Great Danes should be not be significantly different than thinking about the typical residents of Amsterdam, Stockholm, or Copenhagen.

On the other hand, if stereotypes are a distinct form of general semantics, these brain regions should not participate in the retrieval of social knowledge. Functional neuroimaging studies have routinely demonstrated that many social-cognitive tasks recruit a network of brain regions—including the medial prefrontal cortex (MPFC), posterior cingulate, bilateral temporoparietal junction, and anterior temporal cortex—that distinguish them from closely-matched tasks that require participants to engage in nonsocial processing (Mitchell, 2009b). To the extent that stereotypes are indeed a unique form of knowledge, their retrieval may likewise rely on this network. In the current study, we used functional magnetic resonance imaging (fMRI) to arbitrate between these predictions by scanning participants while they alternately answered questions on the basis of their knowledge of social and nonsocial categories.

METHOD

Participants

Nineteen right-handed college undergraduates and community members from the Boston suburbs (9 female, age range 19-28, mean age 22.2 years) with no history of neurological problems participated in exchange for monetary payment. All participants provided informed consent in a manner approved by the Committee on the Use of Human Subjects in Research at Harvard University.

Stimuli and behavioral procedure

During fMRI scanning, participants completed two semantic knowledge tasks. During the *categorical knowledge* task, participants answered a series of questions that required semantic knowledge about categories of people or categories of nonsocial stimuli such as objects. Each trial began with the appearance of two category *labels* (e.g., men, women; guitars, violins).

After 750 ms, a category *feature* appeared below the labels (e.g., watch romantic comedies; have six strings) for an additional 3000 ms. Participants indicated which of the two categories was more likely to have that particular feature by pressing one of two buttons under their left hand. Category labels and features varied between conditions to avoid a reliance on the few trivial features (e.g., size) that can appropriately describe social groups and object categories. Trials were segregated into four functional runs of 40 trials each (20 social and 20 nonsocial). Importantly, social stimuli were rated to be *less* emotionally evocative than nonsocial stimuli—*Ms* (*SDs*) = 4.51 (0.38) vs. 4.65 (0.51)—by a separate group of 57 participants, precluding the possibility that any additional activation associated with social judgments might be due to greater affective processing of social stimuli.

Following the categorical knowledge task, participants also completed one run of a *feature verification* task used to identify the neural regions typically associated with the retrieval of semantic knowledge (Mitchell, Heatherton, & Macrae, 2002). On each of 40 *nonsocial* trials, participants read the name of a fruit (banana, mango) or item of clothing (glove, shirt) and were asked to verify whether an adjective (ripe, threadbare, curious) could be appropriately used to describe the item. On each of 40 *social* trials, participants read the name of a person (John, Mary) and were asked to verify whether the adjective could be used appropriately to describe a person. Adjectives were appropriate and inappropriate descriptors on an equal number of trials, and each trial lasted 4000 ms. To optimize estimation of the event-related fMRI response during both tasks, trials were intermixed in a pseudorandom order and separated by a variable stimulus interval (0 to 10 s; Dale, 1999) during which participants passively viewed a fixation crosshair.

Functional imaging procedure

The experiment was conducted using a 3.0-Tesla Trio scanner with a standard head coil. Functional runs used a gradient-echo, echo-planar pulse sequence (TR = 2000 ms; TE = 35 ms; 3.75 x 3.75 in-plane resolution; 31 axial slices, 5 mm thick; 1 mm skip). Coverage extended to a ventralmost coordinate of z = -22. Stimuli were projected onto a screen that participants viewed by way of a mirror mounted on the head coil. A high-resolution T1-weighted structural scan (MEMPRAGE) was conducted following four runs of the categorical knowledge task (107 volume acquisitions each) and one run of the feature verification task (210 acquisitions).

FMRI data were preprocessed and analyzed using SPM2 (Wellcome Department of Cognitive Neurology, London, UK). First, functional data were time-corrected for differences in acquisition time among slices and realigned to correct for head movement. Functional data were then transformed into a standard anatomical space (3-mm isotropic voxels) based on the ICBM 152 brain template (Montreal Neurological Institute). Normalized data were then spatially smoothed using an 8-mm full-width-at-half-maximum Gaussian kernel. Preprocessed images were analyzed using the general linear model, in which trials were modeled using a canonical hemodynamic response function, its temporal derivative, and additional covariates of no interest (a session mean and a linear trend). Comparisons of interest were implemented as linear contrasts using a random-effects model. A Monte Carlo simulation of our whole-brain volume was used to specify the minimum cluster extent necessary to obtain an experiment-wide statistical criterion of $p < .05$, corrected for multiple comparisons. Additional statistical comparisons between conditions were conducted using ANOVA procedures on the parameter estimates associated with each trial type.

RESULTS

Behavioral data

During the categorical knowledge task, participants responded significantly faster during social (*M* = 1467 ms, *SD* = 165 ms) than nonsocial trials (*M* = 1564 ms, *SD* = 181 ms), *t*(18) = 4.27, *p* < .001, *Cohen's d* = 1.01, making it unlikely that any additional activation associated with social judgments is a result of the greater complexity or difficulty of social stimuli. Item analyses demonstrated that participants converged on the same response equally often for social (*M* = 92%) and nonsocial stimuli (*M* = 89%), *t*(158) = 1.22, *p* = .23, *d* = 0.10. During the feature verification task, participants responded faster during person (*M* = 1063 ms, *SD* = 137) than object trials (*M* = 1133 ms, *SD* = 129), *t*(18) = 3.54, *p* < .002, *d* = 0.83.

Functional imaging data

For the categorical knowledge task, we first used a whole-brain, random-effects analysis to identify cortical regions that were more active during judgment of nonsocial than social categories. Inconsistent with the claim that social knowledge draws on similar processing as other forms of semantic memory, the contrast of *nonsocial* > *social* identified a set of brain regions regularly associated with semantic processing, including left-lateralized inferior frontal gyrus and inferotemporal cortex (Figure 1.1 and Table 1.1). Importantly, social categories elicited no additional response over baseline in both inferior frontal and inferotemporal regions (both *p*s > .14).

These results were confirmed in region-of-interest analyses from the feature verification task. Replicating earlier work (Mitchell, et al., 2002), the comparison of *object* > *person* also identified left-lateralized regions in inferior frontal gyrus and inferotemporal cortex typically

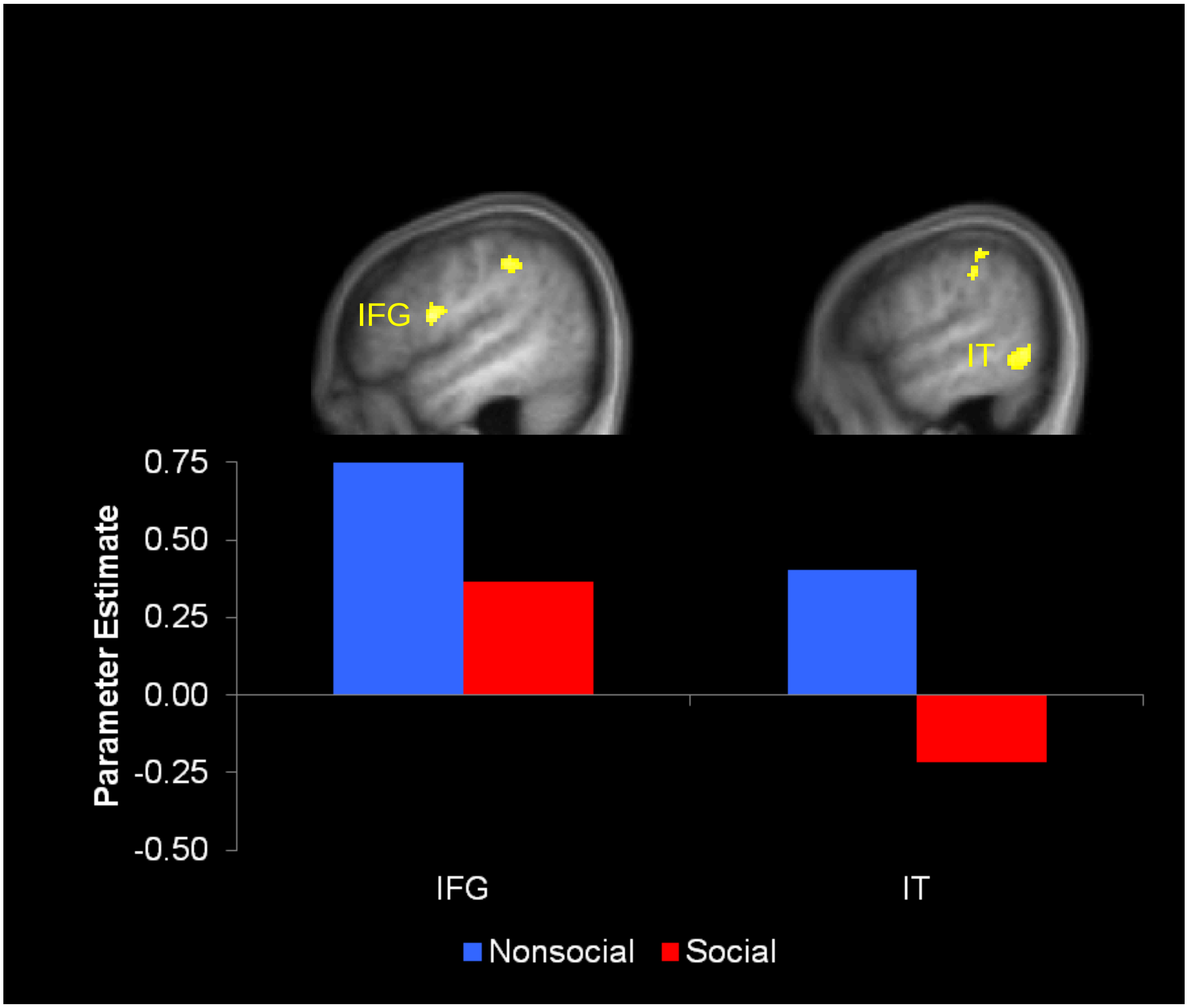

Figure 1.1 Brain regions identified from the contrast of *nonsocial* > *social* for the categorical knowledge task. Whole-brain, random-effects analyses ($p < .05$, corrected) revealed left-lateralized regions of inferior frontal gyrus (IFG) and inferotemporal (IT) cortex that responded robustly during judgments of nonsocial categories, but did not respond differently from baseline during judgments of social categories. Regions are displayed on sagittal images of participants' mean normalized brain (x = -50 and -58, respectively). Bar graphs display the mean parameter estimates from these regions for nonsocial (red) and social (blue) trials.

Table 1.1 Peak voxel and number of voxels for brain regions obtained from the random-effects contrasts of *nonsocial* > *social* trials on the categorical knowledge task and *object* > *person* trials on the feature verification task, *p* < .05, corrected.

Region	*x*	*y*	*z*	*Voxels*	*t*
Categorical Knowledge (Nonsocial > Social)					
Inferior parietal lobule	-46	-36	44	154	4.49
Inferotemporal cortex	-58	-60	-4	130	4.43
Corpus callosum	6	12	22	129	5.50
Inferior frontal gyrus	-48	6	18	95	4.79
Superior frontal gyrus	14	-18	68	95	4.08
Subcentral gyrus	42	-10	24	86	4.79
Middle frontal gyrus	-44	38	12	83	4.36
Superior frontal sulcus	-24	10	56	67	4.81
Feature Verification (Object > Person)					
Inferotemporal cortex	-48	-52	-12	88	4.05
Inferior frontal gyrus	-40	32	14	80	4.66

Note: *T*-tests reflect the statistical difference between the two conditions, as computed by SPM2. Coordinates refer to the stereotaxic space of the Montreal Neurological Institute (MNI).

associated with semantic processing (Table 1.1).

These regions were subsequently interrogated for differences between social and nonsocial trials during the categorical knowledge task. Consistent with the whole-brain analysis, inferior frontal gyrus displayed greater response to nonsocial than social categories, $t(18) = 3.07$, $p < .007$, $d = 0.72$. A marginally significant difference was also observed in inferotemporal cortex, $t(18) = 1.91$, $p = .07$, $d = 0.45$.

These results were confirmed in region-of-interest analyses from the feature verification task. Replicating earlier work (Mitchell, et al., 2002), the comparison of *object* > *person* also identified left-lateralized regions in inferior frontal gyrus and inferotemporal cortex typically associated with semantic processing (Table 1.1). These regions were subsequently interrogated for differences between social and nonsocial trials during the categorical knowledge task. Consistent with the whole-brain analysis, inferior frontal gyrus displayed greater response to nonsocial than social categories, $t(18) = 3.07$, $p < .007$, $d = 0.72$. A marginally significant difference was also observed in inferotemporal cortex, $t(18) = 1.91$, $p = .07$, $d = 0.45$.

We next identified regions in which neural responses were greater for social than nonsocial categorical knowledge with the use of a whole-brain, random-effects analysis of trials on the categorical knowledge task. The contrast of *social* > *nonsocial* identified the network of brain regions previously associated with inferences about mental states: Dorsal and ventral aspects of the MPFC, posterior cingulate, and bilateral temporoparietal junction (see Figure 1.2 and Table 1.2). These results were confirmed by analyses of data from the feature verification task. Replicating earlier work (Mitchell, et al., 2002), the comparison of *person* > *object* also identified MPFC and a left-lateralized region of temporoparietal junction (Table 1.2). These regions were subsequently interrogated for differences between social and nonsocial trials during

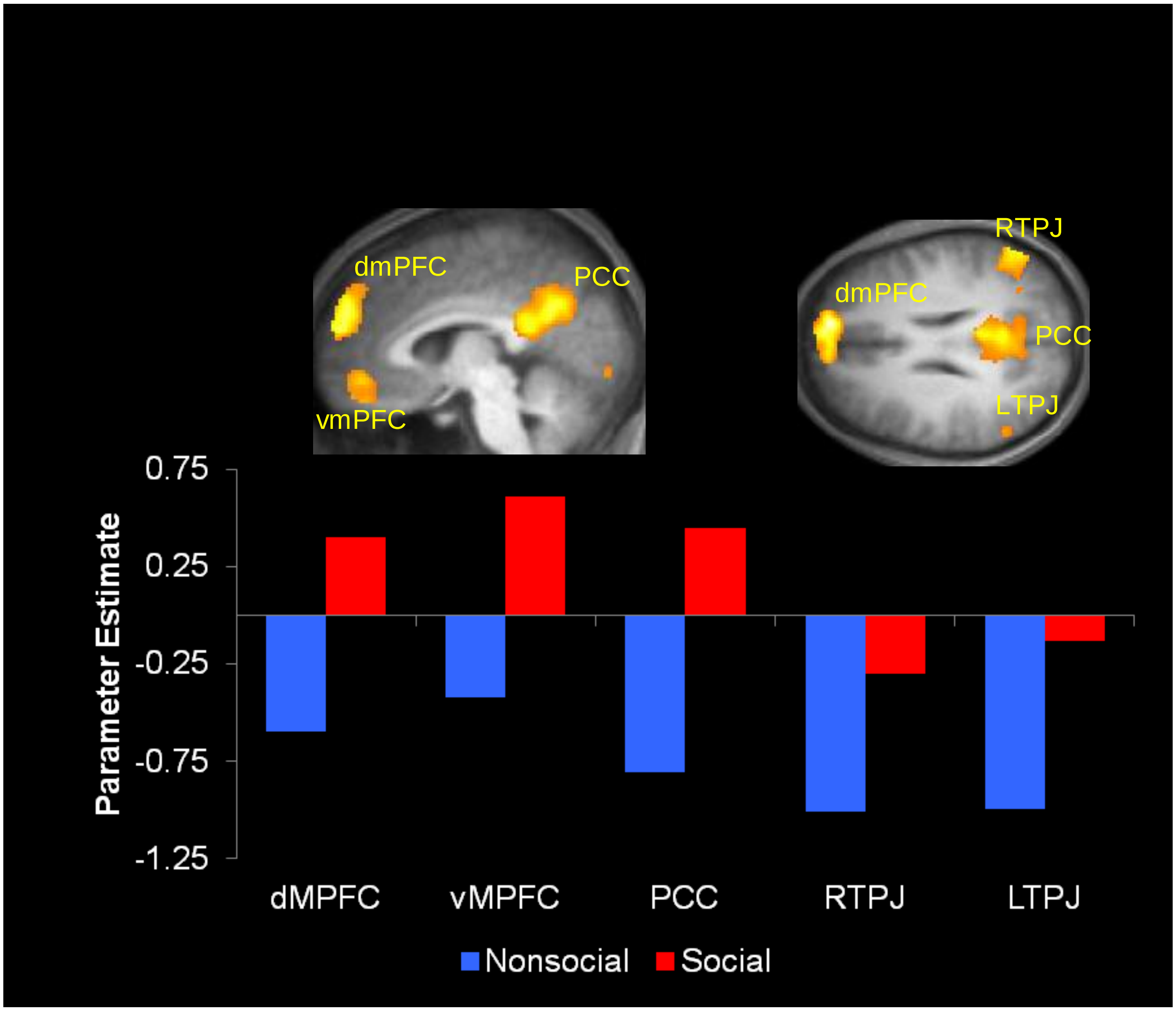

Figure 1.2 Brain regions identified from the contrast of *social* > *nonsocial* for the categorical knowledge task. Whole-brain, random-effects analyses (p < .05, corrected) revealed dorsal and ventral aspects of the MPFC, posterior cingulate cortex (PCC), and left and right temporoparietal junction (TPJ). Regions are displayed on both sagittal (x = -4) and axial (z = 26) images of participants' mean normalized brain. Bar graphs display the mean parameter estimates from these regions for nonsocial (red) and social (blue) trials.

Table 1.2 Peak voxel and number of voxels for brain regions obtained from the random-effects contrasts of *social* > *nonsocial* trials on the categorical knowledge task and *person* > *object* trials on the feature verification task, *p* < .05, corrected.

Region	*x*	*y*	*z*	*Voxels*	*t*
Categorical Knowledge (Social > Nonsocial)					
Posterior cingulate	-4	-58	28	1826	8.34
Medial prefrontal cortex	-8	56	34	1525	10.86
	-4	48	-8	356	6.47
Middle temporal gyrus	-50	-10	-22	1183	10.66
	60	-2	-22	104	5.39
Temporoparietal junction	-56	-60	24	572	7.57
	56	-56	18	78	5.92
Lingual gyrus	-12	-96	-4	548	8.36
Fusiform gyrus	-26	-74	-16	77	5.78
Superior frontal gyrus	-10	38	50	44	5.67
Feature Verification (Person > Object)					
Medial prefrontal cortex	-2	58	22	1121	5.12
Middle temporal gyrus	-56	-4	-24	621	7.32
Lateral orbital gyrus	38	22	-22	272	4.25
Temporoparietal junction	-56	-64	26	197	4.00
Inferior temporal sulcus	54	-12	-32	170	4.23
Posterior orbital gyrus	-40	20	-16	151	4.38
Anterior thalamic nucleus	-4	0	6	135	4.02
Inferior frontal gyrus	44	30	-8	113	4.22
Precentral gyrus	62	18	16	79	3.80

Note: *T*-tests reflect the statistical difference between the two conditions, as computed by SPM2. Coordinates refer to the stereotaxic space of the Montreal Neurological Institute (MNI).

the categorical knowledge task. Consistent with the whole-brain analysis, greater response to social than nonsocial categories was observed in both MPFC (*t*[18] = 6.91, $p < 10^{-5}$, *d* = 1.63) and temporoparietal junction (*t*[18] = 7.66, $p < 10^{-6}$, *d* = 1.81). To confirm that task difficulty did not partially account for these results, we reconditionalized trials based on a median split of each participant's reaction times, resulting in four trial types: *Nonsocial-fast, nonsocial-slow, social-fast*, and *social-slow*. We then interrogated the regions observed in the primary analyses to ascertain whether any demonstrated differences between "fast" and "slow" trials. None of these regions significantly differed by reaction time: Although the PCC demonstrated a non-significant trend toward greater activity during fast than slow trials (*p* = .07), reaction time did not covary with the response in any other reported region (all *p*s > .30).

Finally, the stimulus set included two types of features (actions and physical attributes) used to assess knowledge of each category. For example, in the case of nonsocial categories, "destroy buildings in Kansas" was a possible action associated with tornados and "be blue" was a physical attribute associated with jeans. Likewise, in the case of social categories, "play video games" was an action associated with geeks and "have wide hips" was a physical attribute that describes women more than men. Intriguingly, both ventral MPFC and PCC demonstrated non-significant trends towards greater response for actions than physical attributes for social trials (both *p*s < .07); no other region differentiated significantly between the two types of features (all *p*s > .15).

DISCUSSION

These findings suggest knowledge about the characteristics of social groups bears little resemblance to knowledge about other (nonsocial) categories. When participants made semantic judgments about a variety of nonsocial objects, brain regions traditionally associated with general semantics were engaged, including left inferior frontal gyrus and inferotemporal cortex. In contrast, making similar semantic judgments about groups of people—such as those based on gender, ethnicity, or occupation—failed to engage these regions. Indeed, the response of left inferior frontal gyrus and inferotemporal cortex during social judgments did not differ from baseline: These regions were no more engaged when participants considered the characteristics of social groups than when participants stared at a fixation cross during periods of baseline.

Instead, stereotypes activated a network of brain regions that have been linked regularly to tasks that involve social cognition, including extensive areas of the MPFC, posterior cingulate, bilateral temporoparietal junction, and anterior temporal cortex. For example, these regions have been observed when perceivers infer the beliefs, feelings, or opinions of others (for reviews, see Frith & Frith, 2006; Mitchell, 2009a, 2009b; Saxe, 2006); view objects moving in a way that implies agency (Castelli, Happe, Frith, & Frith, 2000; Wheatley, Milleville, & Martin, 2007); form impressions of people (Mitchell, Cloutier, Banaji, & Macrae, 2006; Mitchell, Macrae, & Banaji, 2004, 2005; Schiller, Freeman, Mitchell, Uleman, & Phelps, 2009); and even when they think about the global characteristics of people as a class (Mitchell, et al., 2002).

Taken together, the current findings suggest a novel way to think about stereotypes, one in which an understanding of social groups may derive less from general semantic knowledge than from our ability to represent the mental states of the members of a group. Many stereotypes about social groups involve inferences about the predilections and dispositions of their members, such as whether men or women prefer watching basketball, Asian-Americans or African-Americans are more likely to play basketball, or middle class or working class individuals are more likely to attend professional basketball games. The regions identified here in the comparison of *social* > *nonsocial* have also been observed when participants make comparable types of inferences about individuals, such as how much a specific person might enjoy watching or playing sports (Jenkins, Macrae, & Mitchell, 2008; Mitchell, Macrae, & Banaji, 2006). Perhaps we deploy a similar set of processes when attributing mental states to social groups as we do to individuals; that is, perhaps we view such groups as mental agents with distinct likes, desires, and proclivities (Brewer & Harasty, 1996; Hamilton & Sherman, 1996).

In this way, the current findings suggest that stereotyping shares more in common with representing mental states than with semantic knowledge of nonsocial categories. In turn, they demonstrate an important facet of the category-specific nature of semantic memory (Caramazza & Shelton, 1998): Namely, knowledge about social categories is not like other forms of semantic knowledge. As such, the present results challenge longstanding claims that stereotypes may be one of many instances of general semantic knowledge.

The present experiment also builds on previous work that observed preferential engagement of the MPFC when participants consider gender stereotypes (Quadflieg et al., 2009). Because participants in this earlier study were explicitly asked to think about what most other people believe about gender roles, it has previously been unclear whether this MPFC activation might be driven by participants' attempts to think about other minds—that is, to mentalize about how another person would answer these questions—rather than stereotyping *per se*. Here, participants were asked simply to judge social and nonsocial attributes on the basis of their own personal semantic knowledge, thus minimizing any explicit demand to consider how other people might answer the same questions.

Previous research has also identified anterior temporal cortex as a region important for representing social knowledge (for reviews, see Olson, Plotzker, & Ezzyat, 2007; Simmons &

Martin, 2009). For example, the anterior temporal cortex has been observed when participants judge words that describe the personality of individual people (Ross & Olson, 2010; Zahn et al., 2009; Zahn et al., 2007b) or encode biographical details about fictional persons (Simmons, Reddish, Bellgowan, & Martin, 2010). Here, we extend this work by demonstrating that the anterior temporal lobe is likewise engaged when drawing on knowledge about social groups: Stereotyping trials were associated with sizeable activations in bilateral portions of middle temporal gyrus that extended rostrally into anterior temporal cortex.

Throughout its history, social psychologists have given a considerable amount of empirical attention to social group dynamics—how we think about groups, how being in a group influences the behavior of individuals, how groups interact with each other, how intergroup conflict and hegemony arise, etc. (for reviews, see Dovidio & Gaertner, 2010; Hackman & Katz, 2010; Yzerbyt & Demoulin, 2010). Although researchers have increasingly brought the methods of cognitive neuroscience to bear on questions of social psychological interest, few such social neuroscience studies have examined how we think about and are influenced by groups. Here, we demonstrate that such emerging methodologies can furnish new insights into the nature of human intergroup cognition, including the current demonstration that knowledge about social categories shares little in common with other forms of semantic knowledge.

STUDY 2

COMMON BRAIN REGIONS WITH DISTINCT PATTERNS OF NEURAL RESPONSES DURING MENTALIZING ABOUT GROUPS AND INDIVIDUALS

Contreras, J. M., Schirmer, J., Banaji, M. R., & Mitchell, J. P. (2013). Common brain regions with distinct patterns of neural responses during mentalizing about groups and individuals. *In press at Journal of Cognitive Neuroscience*.

INTRODUCTION

One of the cognitive abilities that best separates humans from other animals is the capacity to form complex representations of the mental states of others. Humans see a person cry and assume he is sad. They see a person extend her hand toward an object and infer she intends to reach for it. Spontaneously and without difficulty, humans adopt a theory of mind or intentional stance towards other people (Dennett, 1987; Premack & Woodruff, 1978).

Surprisingly, humans attribute similar mental states to groups of people (Jones, 2010). In ordinary speech, they make statements like "Christians *believe* Adam and Eve were real people," "scientists *hope* to understand every aspect of nature," or "companies *think* solely about increasing profits." These expressions cannot refer to the mental states of each group member; surely, some Christians do not believe Adam and Eve were real people, no single scientist hopes to grasp all aspects of nature, and no actual company executive spends every waking hour fantasizing about profits. These expressions refer to the beliefs, thoughts, and desires of groups of people, even though such groups are not conscious entities that can have this kind of internal mental experience. Nonetheless, humans readily use the same mental state vocabulary to describe the "minds" of groups as they typically do to describe the mental states of individuals.

Perhaps claims about the mental states of groups are merely examples of figurative language that make it easier to communicate about collections of individuals. That is, it may be a linguistic convenience to say "senior citizens want the president to stay in office" rather than "88% of citizens 65 and older polled in a recent representative survey would vote for the president if he were to run for reelection." Alternatively, references to the mental states of groups may be legitimate instances of theory of mind. In the same way humans endow objects, nonhuman animals, and fictional entities with complex mental states that they do not possess

(Epley, Waytz, & Cacioppo, 2007), humans may impute legitimate mental states to groups. That is, humans may naturally perceive a group mind.

The idea of a group mind coheres with previous research on stereotypes as group mental states that inform mentalizing about individual group members (Ames, 2005; Ames & Mason, 2012). In this research, perceivers use stereotypes about the groups to which targets belong to infer the intentions and preferences of these targets (Bottom & Paese, 1997; Plous, 1993; Sagar & Schofield, 1980), especially when they view targets to be substantially different from them (Ames, 2004a, 2004b; Ames, Weber, & Zou, 2012). Further evidence of a link between mentalizing and stereotyping is provided by a neuroimaging experiment that suggests that these two operations recruit similar brain regions (Contreras, Banaji, & Mitchell, 2012).

These brain regions—medial prefrontal cortex (MPFC), anterior temporal lobe (ATL), temporoparietal junction (TPJ), and medial parietal cortex—respond robustly across a wide range of situations in which perceivers represent mental states (for reviews, see Frith & Frith, 2006; Gallagher & Frith, 2003; Mitchell, 2009a; Saxe, 2009). For example, these brain regions show increased activity when participants read stories about others' beliefs (Fletcher et al., 1995; Saxe & Kanwisher, 2003), view cartoons that imply mental states (Brunet, Sarfati, Hardy-Baylé, & Decety, 2000; Gallagher et al., 2000), see objects move intentionally (Castelli, et al., 2000; Wheatley, et al., 2007), think about the thoughts of historical characters (Goel, Grafman, Sadato, & Hallett, 1995), and infer the mental states of competitors in strategy games (Gallagher, Jack, Roepstorff, & Frith, 2002; McCabe, Houser, Ryan, Smith, & Trouard, 2001).

If humans adopt a theory of mind about groups, then these brain regions should show robust neural activity when perceivers consider the mental states of groups. Previous research is suggestive: Viewing photographs and videos of social interactions recruits some of the same brain regions that respond to mental state inferences about single individuals (Centelles, Assaiante, Nazarian, Anton, & Schmitz, 2011; Iacoboni et al., 2004; Wagner, Kelley, & Heatherton, 2011). But the participants in these studies were not asked to infer the mental states of groups and no comparison was made between inferences about the mental states of groups and individuals. Therefore, to test the hypothesis that humans adopt a theory of mind about groups, we conducted an experiment in which participants underwent functional magnetic resonance imaging (fMRI) while alternately inferring the mental states of groups and individual group members.

EXPERIMENT 1

METHOD

Participants

Participants were 25 college students and community members from Cambridge, MA (9 male, 16 female; age range: 19-27, *M* = 22.0) who were right-handed, had no history of neurological problems, and provided informed consent in a manner approved by the Committee on the Use of Human Subjects in Research at Harvard University. Three additional participants were excluded: One for excessive head movement (more than 100 instances of at least 1 mm of movement or 1° of rotation from one volume to the next) and two for failing to respond to more than 20% of experimental trials.

Stimuli and behavioral procedure

Participants were scanned using fMRI while performing a *photograph judgment* task in which they viewed 80 photographs that depicted groups of people that ranged in size from 3-180 individuals (*M* = 12.5). On each trial, participants were asked to consider either the group as a whole or a single member of the group. During *group* trials, a blue border appeared around the entire photograph, whereas during *member* trials, the blue border appeared around the face or body of a single person in the group. Participants viewed 40 of the photographs in the member condition, pseudo-randomly selected for every participant from the stimuli set of group photographs

For each photograph, participants were asked to perform one of two tasks. On *mental* trials, participants were oriented toward the mental states of the targets with the cue, “Enjoy a long car ride?”. On group versions of mental trials, participants judged how much the group would enjoy a long car ride together. For member versions of mental trials, participants judged how much the group member would enjoy a long car ride alone. This question was chosen because it could be answered equally well for both groups and individuals, and because group enjoyment depends on the pleasantness of interactions between group members and is not simply reducible to the enjoyment of group members considered individually.

On *physical* trials, participants were oriented towards physical properties of the targets with the cue, “Stay afloat?”. On group versions of physical trials, participants judged how well the group would stay afloat in a life raft. For member versions of physical trials, participants judged how well the group member around would stay afloat in a pair of arm flotation devices. Prior to the start of the experiment, participants saw photographs of these two flotation devices and read descriptions about them: A Boeing life raft six feet in diameter and capable of holding up to 500 pounds and inflatable arm bands that are recommended for children but are capable of holding up to 120 pounds. This question was chosen because it could be answered equally well for both groups and individuals, and because it requires participants to evaluate targets without considering their mental states.

Each trial began with one of the two cue phrases. After 500 ms, the photograph appeared under the cue, along with a four-point Likert scale (1 = Least, 4 = Most). The cue, photograph, and scale remained onscreen for an additional 3250 ms, during which participants indicated their response using a button box in their left hand. For the last 250 ms of a trial, a white fixation cross appeared in the middle of the screen. To optimize estimation of the event-related fMRI response, trials were intermixed in a pseudorandom order and separated by a variable stimulus interval (0-14 s) during which participants passively viewed a fixation crosshair (Dale, 1999). Trials were segregated into four functional runs, each of which consisted of 60 trials (20 trials in each group condition, 10 trials in each member condition).

Functional imaging procedure

Imaging data were acquired on a 3.0 Tesla Siemens Tim Trio scanner (Siemens, Erlangen, Germany) with a standard head coil in the Center for Brain Science at Harvard University. Functional runs used a gradient-echo, echo-planar imaging (EPI) pulse sequence (TR = 2000 ms; TE = 30 ms; flip angle = 85°; field of view = 216 x 216 mm; matrix = 72 x 72; in-plane resolution = 3 x 3 mm; slice thickness = 4 mm). Thirty-one interleaved axial slices parallel to the AC-PC line were obtained to cover the whole cerebrum. The photograph judgment task consisted of 4 runs of 160 volume acquisitions each. Each of the functional runs was preceded by 8 s of gradient and radio frequency pulses that allowed the scanner to reach steady-state magnetization. After the functional runs in each experiment, a high-resolution T1-weighted structural scan (MEMPRAGE) was conducted.

Functional imaging data analysis

FMRI data were preprocessed and analyzed using Statistical Parametric Mapping 8 (SPM8; Wellcome Department of Cognitive Neurology, London, United Kingdom) and in-house MATLAB code (MathWorks, Natick, MA) written by Dylan Wagner (Dartmouth College, Hanover, NH). To correct for head movement, a rigid-body transformation realigned images within each run and across all runs using the first functional image as a reference. Realigned images were unwarped to reduce any additional distortions caused by head movement. Unwarped data were normalized into a stereotaxic space (2-mm isotropic voxels) based on the SPM8 EPI template that conforms to the ICBM 152 brain template space and approximates the Talairach and Tournoux atlas space. Normalized images were spatially smoothed using a Gaussian kernel (8-mm full-width-at-half-maximum) to maximize signal-to-noise ratio and reduce the impact of individual differences in functional neuroanatomy. Finally, individual runs were analyzed on a participant-by-participant basis to find outlier volumes with Artifact Detection Toolbox (ART; McGovern Institute for Brain Research, Cambridge, MA). Outliers were defined as volumes in which participant head movement exceeded 0.5 mm or 1° and volumes in which overall signal were more than three standard deviations outside the mean global signal for the entire run.

For each participant, a general linear model (GLM) was constructed to include task effects and nuisance regressors (run mean, linear trend to account for signal drift over time, six movement parameters computed during realignment, and, if any, outlier scans identified by ART and trials in which participants did not provide a response). To compute unweighted (β) and weighted (t) parameter estimates for each condition at each voxel, the GLM was convolved with a canonical hemodynamic response function (HRF) as well as its temporal and spatial derivatives. These derivatives explain a significant portion of BOLD variability above and beyond the canonical HRF (Henson, Rugg, & Friston, 2001). Trials were modeled as events of durations equal to their respective reaction times to account for differences in RTs across conditions (Grinband, Wager, Lindquist, Ferrera, & Hirsch, 2008).

Comparisons of interest were implemented as linear contrasts. Given the large sample size, significant voxels were identified using a voxel-wise statistical criterion of $p < 10^{-5}$. Regions-of-interest (ROIs), defined using MarsBar (Centre IRMf, Marseille, France) and in-house MATLAB code, were required to exceed 32 voxels in extent, establishing an experiment-wide statistical threshold of $p < .05$, corrected for multiple comparisons, on the basis of Monte Carlo simulations (Slotnick, Moo, Segal, & Hart, 2003). Voxels at the intersection of ROIs from different contrasts were identified using xjView (Stanford University, Palo Alto, CA). Contrasts maps were overlaid on the same anatomical template without recalculation of the statistical thresholds used to generate each contrast map. Therefore, conjunctions use conservative statistical thresholds that can increase our confidence in them. Additional statistical comparisons were conducted in MATLAB using paired-samples *t*-tests on the parameter estimates associated with each trial type.

RESULTS

Behavioral data

Means and standard deviations of responses and response times are displayed in Table 2.1. On mental trials, participants judged that groups (*M* = 3.11) would enjoy a long car ride together less than single members (*M* = 3.39), *t*(24) = 2.84, *p* < .01, *Cohen's d* = 0.58. On physical trials, participants did not judge groups (*M* = 2.79) differently from single members (*M* = 2.84), *t*(24) =

Table 2.1 Participants' responses and response latencies for tasks in Experiments 1 and 2.

Target	*Responses*		*Response Latencies*	
		Experiment 1: Photograph judgment task		
	Mental	*Physical*	*Mental*	*Physical*
Group	3.11^{a} (0.48)	2.79^{b} (0.61)	1.87^{a} (0.35)	1.77^{b} (0.30)
Member	3.39^{c} (0.54)	2.84^{ab} (0.78)	1.89^{a} (0.37)	1.92^{a} (0.35)
		Experiment 2: Photograph judgment task		
	Mental	*Physical*	*Mental*	*Physical*
Group	2.97^{ab} (0.38)	2.88^{ab} (0.51)	1.88^{ab} (0.33)	1.81^{abc} (0.32)
Member	3.03^{ab} (0.31)	2.94^{a} (0.55)	1.88^{a} (0.33)	1.87^{a} (0.21)
Individual	3.50^{c} (0.38)	3.05^{b} (0.59)	1.79^{bc} (0.31)	1.73^{c} (0.37)
		Experiment 2: False belief localizer		
	Belief	*Photo*	*Belief*	*Photo*
--	$.93^{\mathbf{a}}$ (.10)	$.87^{b}$ (.13)	3.06^{a} (1.05)	3.29^{b} (0.85)

Note: Means and, in parentheses, standard deviations. In the photograph judgment task, participants' responses refer to their judgments on a four-point Likert scale (1 = Least, 4 = Most). In the false belief localizer, participants' responses refer to their proportion of correct responses to the questions about the stories. Response times are displayed in seconds. For each dependent variable, means sharing a superscript do not differ significantly at $p < .05$, as computed in paired-samples t-tests.

0.29, $p = .78$, $d = 0.06$. Response latencies were closely matched for mental judgments in the group ($M = 1.87$) and member ($M = 1.89$) conditions, $t(24) = 0.87$, $p = .39$, $d = 0.18$, but participants responded more quickly to physical judgments about groups ($M = 1.77$) than single members ($M = 1.92$), $t(24) = 3.62$, $p < .01$, $d = 0.74$.

The behavioral data suggest that participants did not answer questions about the mental states of groups by inferring the mental state a single group member. Inferring the mental state of a single member rather the group as a whole would mean that participants need not take the size of the group into account in their inference. However, responses in group trials correlated with group size: The larger the group, the less enjoyable the long car ride, Fisher-transformed $r(24) = -0.17$, $p < 10^{-6}$.

Functional imaging data

Trials from the photo judgment task were conditionalized on the basis of question (mental, physical) and target (group, member), resulting in 4 conditions of interest. A whole-brain contrast identified voxels in which BOLD activity was greater during trials in which participants answered questions about the mental states of groups than their physical appearance (*group mental* > *group physical*). This contrast identified ventral and dorsal aspects of MPFC, bilateral ATL, bilateral TPJ, and medial parietal cortex (Table 2.2). Consistent with earlier research, a similar whole-brain contrast of single member trials (*member mental* > *member physical*) identified the same brain regions (Table 2.2). For a formal test that these two contrasts identified overlapping brain regions, we identified voxels in the intersection of the ROIs defined by each contrast. This analysis identified voxels in dorsal and ventral MPFC, bilateral ATL, bilateral TPJ, and medial parietal cortex (Figure 2.1).

Table 2.2 Brain regions identified in whole-brain, random-effects contrasts (*group mental* > *group physical*, *member mental* > *member physical*) in the photograph judgment task of Experiment 1, *p* < .05, corrected for multiple comparisons.

	Group Mental > Group Physical				*Member Mental > Member Physical*			
Region	*x*	*y*	*z*	*k*	*x*	*y*	*z*	*k*
Temporoparietal junction								
Right	54	-51	16	1218	56	-57	14	891
Left	-46	-77	30	811	-42	-63	28	1471
Medial prefrontal cortex								
Ventral	-4	53	-10	1075	-4	47	-8	579
Dorsal	-8	61	36	749	-6	45	50	490
Anterior temporal cortex								
Left	-56	-9	-22	798	-54	-11	-16	1133
Right	56	-1	-22	769	42	19	-32	897
Medial parietal cortex	-8	-55	16	524	0	-49	28	95

Note: From left to right, columns list the names of regions obtained from whole-brain, random-effects contrasts, the stereotaxic Montreal Neurological Institute coordinates of their peak voxels, and their size in number of voxels (*k*).

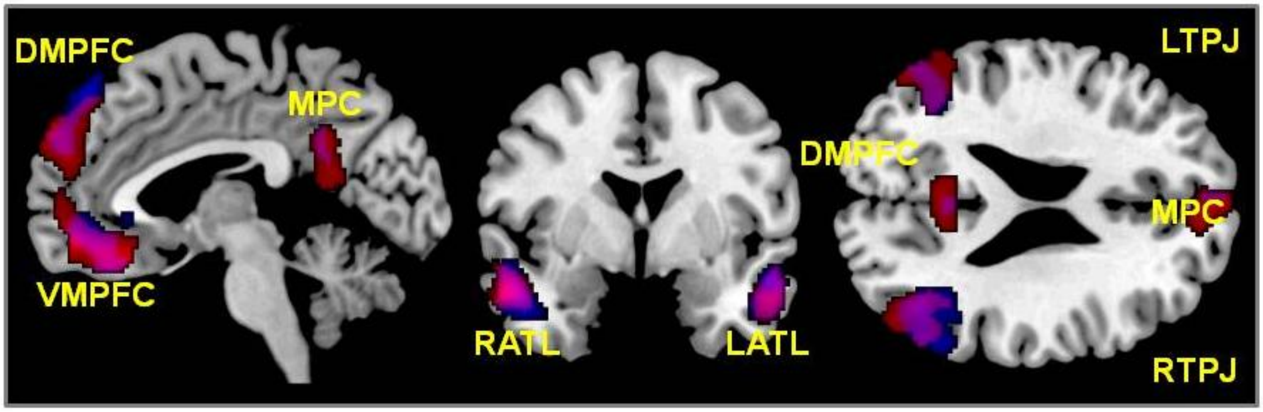

Figure 2.1 T-maps computed by whole-brain, random-effects contrasts *group mental* > *group physical* (blue) and *member mental* > *member physical* (red) in the photo judgment task of Experiment 1. Voxels at the intersection of the two *T*-maps are displayed in purple. *T*-maps are displayed in sagittal (x = -4), coronal (y = 0), and axial (z = 24) slices of a high-resolution anatomical template. These contrasts identified ventromedial prefrontal cortex (VMPFC), dorsomedial prefrontal cortex (DMPFC), bilateral anterior temporal lobe (RATL and LATL), bilateral temporoparietal junction (RTPJ and LTPJ), and medial parietal cortex (MPC).

As an additional test of the hypothesis that brain regions recruited by mental state inferences about individuals are also recruited by mental state inferences about groups, the parameter estimates of the group contrast (*group mental* > *group physical*) were extracted for the brain regions defined by the member contrast (*member mental* > *member physical*), averaging over all voxels in each ROI (Table 2.3). To test that inferences about the mental states of groups also recruit these brain regions, we examined whether BOLD activity in these parameter estimates were significantly higher than zero. This was the case for every ROI, all ts(24) > 5.00, all ps < 10^{-4}, suggesting that perceivers also recruit these regions when they represent the mental states of groups.

Moreover, to test whether inferences about the mental states of groups recruit these brain regions as robustly as inferences about the mental states of single members, we examined whether differences in BOLD activity between the parameter estimates of the group and member condition (*member mental* > *member physical*) were statistically equivalent (Table 2.3). Although ventral MPFC showed a marginally larger response to groups than single members, $t(24) = 1.84$, $p = .08$, every other ROI showed equivalent responses to mental state inferences about groups and single members, all ts(24) < 1.60, all ps > .12, suggesting that perceivers recruit these brain regions as strongly when they represent the mental states of groups as when they perform similar inferences about individuals.

DISCUSSION

We hypothesized that, if humans adopt a theory of mind about groups, then brain regions engaged by theory of mind about individuals—MPFC, bilateral ATL, bilateral TPJ, and medial parietal cortex—should show robust neural activity when perceivers consider the mental states of

Table 2.3 Mean parameter estimates of the contrasts *group mental* > *group physical* and *member mental* > *member physical*, extracted from brain regions identified in whole-brain, random-effects contrasts in the photograph judgment task of Experiment 1 (*member mental* > *member physical*) and Experiment 2 (*individual mental* > *individual physical*).

	Member Mental > Member Physical		*Individual Mental > Individual Physical*		
Region	β_{group}	β_{member}	β_{group}	β_{member}	*d*
Temporoparietal junction					
Right	0.55 (0.09)	0.39 (0.08)	0.44 (0.09)	0.39 (0.08)	0.13
Left	0.48 (0.09)	0.45 (0.08)	0.40 (0.11)	0.46 (0.12)	0.11
Medial prefrontal cortex					
Ventral	0.83 (0.13)	0.50 (0.11)	0.47 (0.08)	0.23 (0.12)	0.47
Dorsal	1.00 (0.19)	0.78 (0.15)	0.35 (0.10)	0.34 (0.08)	0.03
Anterior temporal cortex					
Left	0.37 (0.06)	0.37 (0.05)	0.25 (0.05)	0.23 (0.05)	0.08
Right	0.41 (0.06)	0.35 (0.05)	0.32 (0.07)	0.26 (0.06)	0.19
Medial parietal cortex	0.49 (0.10)	0.30 (0.08)	0.25 (0.08)	0.19 (0.07)	0.15

Note: From left to right, columns list the names of regions obtained from whole-brain, random-effects contrasts and the parameter estimates of the group (*group mental* > *group physical*) and member (*member mental* > *member physical*) contrasts, averaging over all voxels in each region. The rightmost column lists the effect size (Cohen's *d*) of the difference between parameter estimates of the group and member contrasts in Experiment 2. Numbers in parentheses denote standard errors. Parameter estimates are reliably larger than zero in every region, all *p*s < .01. Group and member mean parameters do not differ reliably in any region, all *p*s > .05.

groups. In support of this hypothesis, these brain regions were more responsive to inferences about mental states of both groups and individuals than to inferences about the physical appearance of both groups and individuals.

As such, the results of the present experiment suggest that common brain regions respond during inferences about mental states of groups and individuals. But this result is puzzling given that people do not consider the minds of single individuals and the "minds" of groups to be equivalent. For instance, people are more likely to attribute intentions than emotions to groups, but they do not show such a bias in imputing mental states to individuals (Knobe & Prinz, 2008). Therefore, though common brain regions may respond preferentially to theory of mind about individuals and groups, they may have distinct representations of mental state inferences about groups and mental state inferences about individuals. To test this hypothesis, we turned to multivoxel pattern analysis (MVPA), which takes as its unit of analysis not the responses of individual voxels but the responses of multiple voxels or *multivoxel patterns* (for review, see Weil & Rees, 2010). Differences in multivoxel patterns across stimuli index differences in the underlying representations of these stimuli. For example, face and house percepts elicit distinct multivoxel patterns in ventral temporal cortex, suggesting that the visual system contains distinct representations of faces and houses (Haxby et al., 2001). In the present experiment, we identified brain regions in which multivoxel patterns differentiate mental state inferences about groups from similar inferences about individual group members. In other words, this analysis aimed to identify brain regions that contain distinct representations of mental state inferences about groups and individuals.

We conducted this analysis in an experiment in which participants performed the same photograph judgment task from Experiment 1. But whereas participants in Experiment 1 only viewed some of the photographs in their member condition, participants in Experiment 2 viewed all photographs in their group and member condition. The larger number of trials in the member condition as well as the equal number of trials in the group and member conditions allows the use of MVPA. Specifically, this modification ensures that the parameter estimates in the multivoxel patterns of the group and member conditions are equal in statistical power and reliability. To ascertain that we examined the brain regions that have been previously identified as responding preferentially to mentalizing, this version of the task included an additional set of trials that required participants to make mental state inferences about individuals outside of a group context. These trials were used as a functional localizer of brain regions engaged by mentalizing. Additionally, participants completed a false belief task (Saxe & Kanwisher, 2003), which provided an independent functional localizer of brain regions engaged by mentalizing.

EXPERIMENT 2

METHOD

Participants

Participants were 14 individuals (8 male, 6 female; age range: 18-23, *M* = 19.9) sampled from the same population as Experiment 1. One participant withdrew from the experiment after becoming ill during data collection.

Stimuli and behavioral procedure

Participants were scanned using fMRI while performing the same photograph judgment task with two modifications. First, whereas participants in Experiment 1 viewed 40 of the photographs in the member condition, participants in Experiment 2 viewed all 80 photographs in their member version. Second, the task included photographs that depicted a single *individual* (40 photographs). These photographs, drawn from the same database, depicted individual people in a variety of settings. Each photograph was presented twice. In one presentation, participants judged how much the individual would enjoy a long car ride alone in one set of trials. In the other presentation, they judged how well the group member would stay afloat in a pair of arm flotation devices. Trials were segregated into eight functional runs, each of which consisted of 50 trials (10 trials in each group and member condition, 5 trials in each individual condition).

Participants in Experiment 2 also completed two runs of a *false belief* localizer used to identify brain regions that represent mental states (Saxe & Kanwisher, 2003). During this task, participants read 12 vignettes that referred to the false belief of a person (*belief* blocks). For example, one belief vignette read, "Jenny put her chocolate away in the cupboard. Then she went outside. Alan moved the chocolate from the cupboard into the fridge. Half an hour later, Jenny came back inside." The corresponding cue read, "Jenny expects to find her chocolate in the:" and the answer choices read "fridge" and "cupboard." Participants also read 12 vignettes that referred to an outdated physical representation, such as a photograph (*photo* blocks). These vignettes had the same logical structure as belief stories, but referred to outdated physical, rather than mental, representations (Zaitchik, 1990). For example, one physical vignette read, "A photograph was taken of an apple hanging on a tree branch. The film took half an hour to develop. In the meantime, a strong wind blew the apple to the ground." The corresponding cue read, "The developed photograph shows the apple on the:" and the corresponding answer choices read "ground" and "branch." Each vignette was presented for 10 s, after which it was replaced by the cue sentence. Participants had 6 s to respond to the cue before the end of the story. Each cue was followed by a fixation crosshair lasting 10 s. Stories were segregated into two functional runs, each of which consisted of 6 belief and 6 photo stories, intermixed in pseudorandom order.

Functional imaging procedure

The functional imaging procedures were identical to those of Experiment 1. However, the photograph judgment task consisted of 8 runs of 130 volume acquisitions each. The false belief localizer consisted of 2 runs of 174 volume acquisitions each. Significant voxels in the photograph judgment task were identified using a voxel-wise statistical criterion of $p < .005$ and ROIs were required to exceed 59 voxels in extent. Given the block design of the false belief localizer, the thresholds in the analysis of this data were $p < 10^{-5}$ and $k = 32$. These thresholds established an experiment-wide statistical threshold of $p < .05$, corrected for multiple comparisons, on the basis of Monte Carlo simulations (Slotnick, et al., 2003).

Functional imaging data analysis

The functional imaging data analysis was identical to that of Experiment 1, with one exception. Information-based functional brain mapping with a multivariate spherical searchlight was conducted for each participant (Kriegeskorte, Goebel, & Bandettini, 2006). For this analysis, fMRI data from the photo judgment task were preprocessed as before, but they were spatially smoothed with a smaller Gaussian kernel (4-mm full-width-at-half-maximum). GLMs were also constructed as before, but each group and member condition was split into one condition with trials from odd runs and another condition with trials from even runs (e.g., *group mental odd*, *group mental even*). Contrast images were created by comparing the parameter estimates of each mental condition to those of its corresponding physical condition in a linear contrast (e.g., *group* *mental odd* > *group physical odd*). Following Misaki, Kim, Bandettini, and Kriegeskorte (2010), the multivariate searchlight analysis used the resulting contrast images (*group mental odd* > *group physical odd*, *group mental even* > *group physical even*, *member mental odd* > *member physical odd*, *member mental even* > *member physical even*) because they reduce the influence of noisy voxels.

For each voxel, we extracted the parameter estimates of each contrast within a spherical neighborhood (8-mm radius; neighborhood size in resampled voxels, *M* = 248, *SD* = 21) similar in shape to those used by Kriegeskorte and colleagues (2006). As such, each neighborhood was associated with four vectors, one for each contrast. These vectors were correlated in four different ways: Two *same*-target correlations (*group odd with group even*, *member odd with member even*) and two *different*-target correlations (*group odd with member even*, *member odd with group even*). These correlations were Fisher-transformed to *z*-values ($z = \frac{1}{2} * \ln\left(\frac{1+r}{1-r}\right)$) and, then, averaged to yield a single *same*-target correlation and a single *different*-target correlation. For each neighborhood, we subtracted the average same-target correlation from the average different-target correlation and assigned the difference to the center voxel. This analysis yielded a correlation difference map expressed in *z*-scores for each participant, indexing the degree to which each voxel exists in a neighborhood in which group trials correlate with each other more than with single member trials and, vice versa, single member trials correlate more with each other than group trials. Finally, a univariate, random-effects analysis identified brain regions that showed higher same- than different-target correlations across participants. For each voxel, we performed a right-tailed one-sample *t*-test against zero with the corresponding *z*-values from all participants (*same* > *different*).

RESULTS

Behavioral data

Means and standard deviations of responses and response times are displayed in Table 2.1. On mental trials, participants indicated that groups (*M* = 2.97) would enjoy a long car ride together as much as single members (*M* = 3.03), *t*(13) = 0.90, *p* = .39, *d* = 0.25. On physical trials, participants did not judge groups (*M* = 2.88) differently from single members (*M* = 2.94), *t*(13) = 0.28, *p* = .78, *d* = 0.08. Response latencies were closely matched for mental judgments in the group (*M* = 1.88) and member (*M* = 1.88) conditions, *t*(13) = 0.19, *p* = .85, *d* = 0.05, as well as for physical judgments about groups (*M* = 1.81) and single members (*M* = 1.89), *t*(13) = 1.41, *p* = .18, *d* = 0.39.

The behavioral data suggest that participants did not answer questions about the mental states of groups by inferring the mental state a single group member. Inferring the mental state of a single member rather the group as a whole would mean that participants need not take the size of the group into account in their inference. However, responses in group trials correlated with group size: The larger the group, the less enjoyable the long car ride, Fisher-transformed *r*(13) = -0.17, $p < 10^{-6}$.

Functional imaging data

Univariate analysis. Trials from the photo judgment task were conditionalized on the basis of question (mental, physical) and target (group, member, individual), resulting in 6 conditions of interest. A whole-brain contrast identified voxels in which BOLD activity was greater during trials in which participants answered questions about the mental states of individuals than their physical appearance (*individual mental* > *individual physical*). Consistent with earlier research, this contrast yielded a set of brain regions that included dorsal and ventral MPFC, bilateral ATL, bilateral TPJ, and medial parietal cortex (Table 2.4).

The parameter estimates of the group and member conditions (*group mental* > *group physical*, *member mental* > *member physical*) were extracted for each of these brain regions, averaging over all voxels in each ROI (Table 2.3). To test that inferences about the mental states of groups recruit these brain regions, we examined whether BOLD activity in the parameter estimates of the group conditions were significantly higher than zero. This was the case for every ROI, all *t*s(13) > 2.98, all *p*s < .01, suggesting that people also recruit these regions when they represent the mental states of groups. To test that inferences about the mental states of groups recruit these brain regions as robustly as inferences about the mental states of single members, we examined whether differences in BOLD activity between the group and member parameter estimates were statistically unreliable. This was the case for every ROI, all *t*s(13) < 1.70, all *p*s > .11, suggesting that people recruit these brain regions as strongly when they represent the mental states of groups as when they perform similar inferences about individuals.

These results were replicated when ROIs were defined by the false belief localizer. A whole-brain contrast identified voxels in which BOLD activity was greater during blocks in which participants read and answered questions about belief stories than closely-matched stories about physical representations (*belief* > *photo*). Consistent with earlier research, this contrast yielded a set of brain regions that included dorsal MPFC, right ATL, bilateral TPJ, and medial parietal cortex (Table 2.4).

The parameter estimates of the group and member conditions (i.e., *group mental* > *group physical*, *member mental* > *member physical*) were extracted for each of these brain regions, averaging over all voxels in each ROI. To test that inferences about the mental states of groups

Table 2.4 Brain regions identified in whole-brain, random-effects contrasts in the photograph judgment task (*individual mental* > *individual physical*) and the false belief localizer (*belief* > *photo*) of Experiment 2, $p < .05$, corrected for multiple comparisons.

	Individual Mental > Individual Physical				*Belief > Photo*			
Region	*x*	*y*	*z*	*k*	*x*	*y*	*z*	*k*
Medial prefrontal cortex								
Dorsal	-8	45	52	1221	8	61	26	699
Ventral	-8	55	2	807	--	--	--	--
Medial parietal cortex	-2	-51	22	860	10	-51	34	2491
Anterior temporal cortex								
Left	-48	-17	-24	813	--	--	--	--
Right	62	-7	-18	123	54	-3	-22	1484
Temporoparietal junction								
Left	-54	-67	32	166	-60	-59	20	561
Right	66	-49	26	80	58	-53	30	888
Cerebellum								
Right	30	-83	-36	376	--	--	--	--
Left	-26	-79	-36	99	-24	-77	-38	333
Superior temporal sulcus	44	-45	6	263	--	--	--	--
Parahippocampal gyrus	-34	-31	-18	130	--	--	--	--
Inferior frontal gyrus (orbital)	-48	29	-14	116	--	--	--	--

Note: From left to right, columns list the names of regions obtained from whole-brain, random-effects contrasts, the stereotaxic Montreal Neurological Institute coordinates of their peak voxels, and their size in number of voxels (*k*).

recruit these brain regions, we examined whether BOLD activity in the parameter estimates of the group conditions were significantly higher than zero. This was the case for dorsal MPFC, right ATL, and bilateral TPJ, all *t*s(13) > 2.18, all *p*s < .03, suggesting that people also recruit these regions when they represent the mental states of groups. Medial parietal cortex showed a similar effect, but the difference was statistically reliable at a marginal level, *t*(13) = 1.60, *p* = .06. To test that inferences about the mental states of groups recruit these brain regions as robustly as inferences about the mental states of single members, we examined whether differences in BOLD activity between the group and member parameter estimates were statistically unreliable. This was the case for every ROI, all *t*s(13) < 1.30, all *p*s > .21, suggesting that people recruit these regions as strongly when they represent the mental states of groups as when they perform similar inferences about individuals.

Multivariate searchlight analysis. A whole-brain contrast identified voxels in neighborhoods in which multivoxel correlations were higher for inferences about the mental states of the same class of targets (e.g., *group odd with group even*) than different classes of targets (e.g., *group odd with member even*). This *same* > *different* contrast yielded a set of brain regions that included dorsal MPFC, medial parietal cortex (precuneus and posterior cingulate cortex), left ATL, and left TPJ (Table 2.5, Figure 2.2). In these brain regions, the patterns of BOLD activity associated with inferences about the mental states of groups was more similar to the pattern for other groups than it was for individual members; likewise, the pattern of BOLD activity associated with inferences about the mental states of individuals was more similar to other individuals than to groups. The *z*-values that correspond to the differences between same- and different-target correlations were reliably larger than zero, all *t*s(13) > 4.46, all *p*s < 10^{-4}. Moreover, the

Table 2.5 Brain regions identified in whole-brain, random-effects contrast *same* > *different* from the multivariate searchlight analysis in Experiment 2, *p* < .05, corrected for multiple comparisons.

Region	*x*	*y*	*z*	*k*	*t*
Dorsomedial prefrontal cortex	-6	41	32	133	3.78
Medial parietal cortex					
Precuneus	-18	-69	54	228	3.78
Posterior cingulate cortex	-8	-49	16	282	3.68
Middle frontal gyrus	-30	33	26	342	3.73
Left anterior temporal cortex	-46	9	-34	282	3.58
Postcentral gyrus	40	-27	54	142	3.58
Lateral superior frontal gyrus	22	5	56	775	3.56
	-18	1	66	165	3.53
Supramarginal gyrus	56	-31	34	166	3.56
Inferior frontal gyrus (orbital)	-34	27	-14	250	3.53
Left temporoparietal junction	-52	-39	24	133	3.44

Note: From left to right, columns list the names of regions obtained from whole-brain, random-effects contrast, the stereotaxic Montreal Neurological Institute coordinates of their peak voxels, their size in number of voxels (*k*), and their mean weighted parameter estimate (*t*).

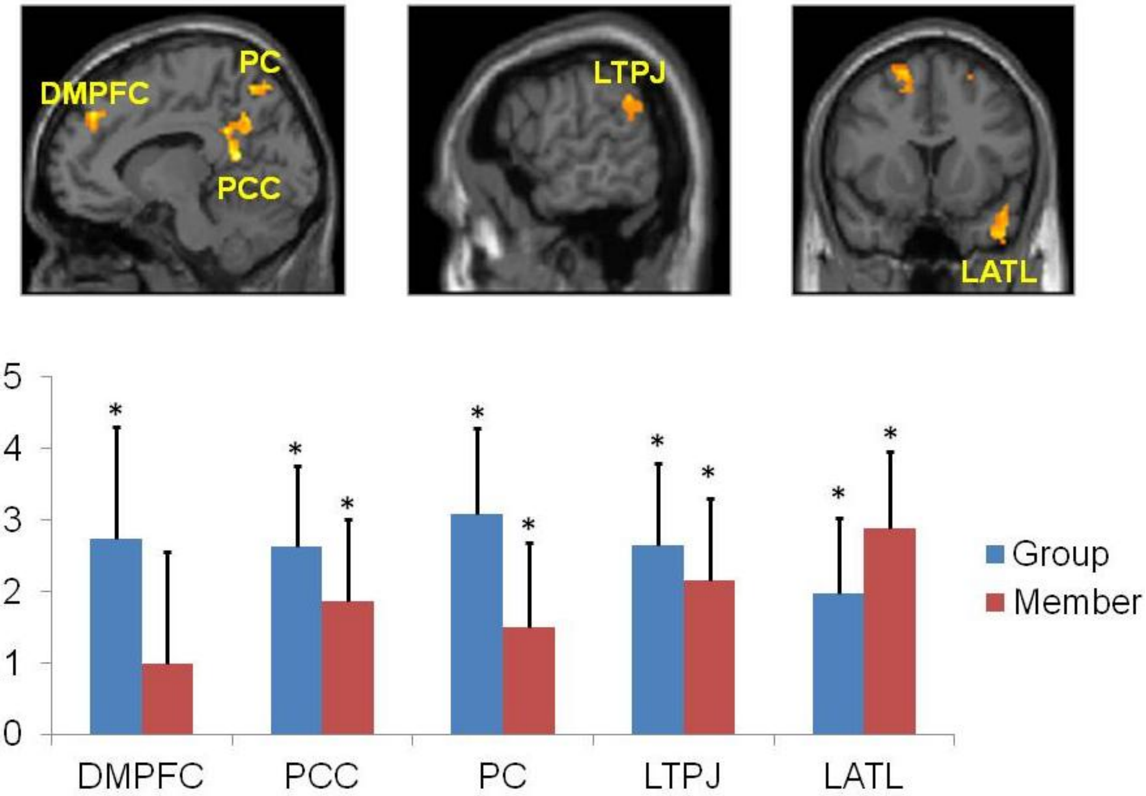

Figure 2.2 T-map computed by the whole-brain, random-effects contrast *same* > *different* from the multivariate searchlight analysis in Experiment 2 and displayed in sagittal (x = -10, x = -60) and axial (y = 13) slices of a high-resolution anatomical template. The analysis identified dorsal medial prefrontal cortex (DMPFC), two clusters in medial parietal cortex: precuneus (PC) and posterior cingulate cortex (PCC), left anterior temporal lobe (LATL), and left temporoparietal junction (LTPJ). Bar graphs display the mean differences of Fisher-transformed correlation averages for group trials (*group odd with group even* > *group odd with member even*, *member odd with group even*) and member trials (*member odd with member even* > *group odd with member even*, *member odd with group even*). Error bars represent 95% confidence intervals in within-subject comparisons (Masson & Loftus, 2003).

magnitude of these differences did not vary across ROIs, all *t*s < 1.65, all *p*s > .12, suggesting that these brain regions were equally sensitive in their discrimination of groups and single members during inferences about mental states.

To ensure that these results were not caused by the group or the member condition exclusively, two additional multivariate searchlight analyses were conducted. A group analysis examined multivoxel correlations in the group condition (*group odd with group even*), whereas a member analysis examined multivoxel correlations in the member condition (*member odd with member even*). By analyzing the two same-target correlations separately, we can examine a correlation difference that detects the presence of a group representation and a correlation that detects the presence of a member representation. Each of these same-target correlations was independently contrasted against the average different-target correlation (*group odd with member even, member odd with group even*).

The correlation differences of the group analysis and the member analysis were extracted separately from the ROIs defined by the original multivariate searchlight analysis that averaged these two correlation differences. If these brain regions contain distinct group and member representations, then we should expect each of these two correlation differences to be reliably larger than zero in each ROI. This was the case in most ROIs (Figure 2.2). The correlation differences from the group analysis were reliably larger than zero in all ROIs, all *t*s(13) > 3.09, all *p*s < 10^{-3}, suggesting that these brain regions contain distinct group representations. In the member analysis, the correlation differences were reliably larger than zero in medial parietal cortex (precuneus and posterior cingulate cortex), left ATL, and left TPJ, all *t*s(13) > 2.84, all *p*s < 10^{-3}, but not in dorsal MPFC, *t*(13) > 1.23, *p* = .12; however, the correlation difference in dorsal MPFC had the predicted direction. These results suggest that these ROIs contain distinct member representations. Together, the group analysis and the member analysis suggest that these brain regions, with the exception of dorsal MPFC, have distinct representations of groups and individuals.

The multivoxel patterns of these brain regions discriminate inferences about the mental states of groups from similar inferences about single members. To determine whether these brain regions discriminated between groups and single members with their univariate responses, parameter estimates from the contrasts (*group mental odd* > *group physical odd*, *group mental even* > *group physical even*, *member mental odd* > *member physical odd*, *member mental even* > *member physical even*) were extracted for each of these brain regions, averaging over all voxels in each ROI. Parameter estimates from odd and even runs of the group contrasts were averaged as were the parameter estimates from odd and even runs of the member contrasts. The differences between these two averages were not statistically reliable in any of the ROIs, all *t*s(13) < 1.53, all *p*s > .15, suggesting that the univariate, unlike the multivariate, responses of these brain regions do not carry information about the target of the mental state inference.

Finally, an additional multivariate searchlight analysis was conducted to ensure that the results of the multivariate searchlight analysis were specific to the group-member distinction. Photographs were pseudo-randomly assigned to two sets (X and Y) such that group and member versions of each photograph were always in the same set. Pseudo-random assignments varied across participant. Experimental trials were reassigned from old group-member conditions to new X-Y conditions (e.g., if photograph 3 is in set X, then its group and member mental trials in the first run are in condition *X mental odd*). Critically, these new conditions contain an equal number of group and member trials. They were used to build new contrasts (*X mental odd* > *X physical odd, X mental even* > *X physical even, Y mental odd* > *Y physical odd, Y mental even* >

Y physical even), which were submitted to a whole-brain multivariate searchlight analysis. At each neighborhood, the average of two different-set correlations (*X odd with Y even*, *X even with Y odd*) was subtracted from the average of two same-set correlations (*X odd with X even*, *Y odd with Y even*) and assigned to the center voxel. A whole-brain random-effects analysis to identify brain regions that showed greater same- than different-set correlations failed to dorsal MPFC, medial parietal cortex (posterior cingulate cortex or precuneus), left ATL, or left TPJ. Instead, it yielded a brain region in left posterior superior temporal gyrus ([*x y z*] = -50, -47, 8).

GENERAL DISCUSSION

People know that groups do not have minds, but nevertheless speak about groups as if they had mental states like those of individuals. In the present experiments, we observed that brain regions that respond preferentially during inferences about the mental states of individuals also responded robustly during inferences about the "mental states" of groups. Irrespective of how they were identified, dorsal and ventral MPFC, bilateral ATL, bilateral TPJ, and medial parietal cortex were more strongly engaged by inferences about the mental states of groups than by similar inferences about physical aspects of these groups. Moreover, these brain regions responded as robustly to inferences about mental states of groups as they did to similar inferences about individual group members. As a whole, these results suggest that not only do humans adopt a theory of mind about groups, but that it is just as robust as is theory of mind about individuals. As such, the results of the present experiment suggest that common brain regions underlie inferences about mental states of groups and individuals.

However, the fact that these brain regions are involved in representing the mental states of groups and individuals leaves open the question of whether these brain regions represent inferences about the mental states of groups and individuals similarly or differently. To determine whether these brain regions contain such target-specific representations (group vs. individual), we carried out a multivariate searchlight analysis to find brain regions that use multivoxel patterns to discriminate mental states inferences about groups from similar inferences about individuals. This analysis identified dorsal MPFC, medial parietal cortex (posterior cingulate cortex and precuneus), left ATL, and left TPJ as repositories of target identity in mental state inferences. These brain regions differentiated between inferences about the mental states of groups and single members despite the fact that the two conditions were closely matched in behavioral responses, response times, information on screen, and univariate BOLD activity. Thus, although the same brain regions represent the mental states of groups and individuals, most of them carry information about whether the target of the mental state inference is a group or an individual.

One could argue that mental state inferences about groups and individuals recruit the same brain regions in the present study because participants inferred the mental state of a single, representative group member when they were asked to mentalize a group. However, two aspects of the data weigh against this possibility. First, multivoxel patterns in most of these brain regions differentiate group and member trials. If participants were answering questions about an individual member in every trial, then multivoxel patterns would not differentiate between group and member trials. The MVPA suggests that these brain regions have distinct representations of groups and individuals. Second, participants indicated that larger groups would enjoy car rides less. If participants had inferred the mental state of a single individual, then their inferences would have been unaffected by group size. For these two reasons, it is unlikely that participants inferred the mental state of a single, representative group member when they were asked to mentalize a group.

In previous research, the brain regions that people recruit in inferences about the mental states of other humans were also engaged by mentalizing about nonhumans, such as objects (Castelli, et al., 2000; Wheatley, et al., 2007), robots (Krach et al., 2008), or other animals (Mitchell, Banaji, & Macrae, 2005). The present findings extend this body of work by showing that these brain regions can also respond robustly to inferences about the mental states of groups of people. Together, these experiments suggest that theory of mind is a cognitive module that can be flexibly deployed for understanding a diverse array of agents (Epley, et al., 2007). Although it is likely that theory of mind evolved to understand the behavior of individual humans (Povinelli & Preuss, 1995), it may have been co-opted by the human mind to increase its understanding of the behavior of nonhuman agents and groups of people.

In sum, the present experiments contribute to our emerging understanding of how the human brain represents information about social groups. Although cognitive neuroscientists have started to document the neural basis of the perception of individuals from different social groups (for reviews, see Cunningham & Van Bavel, 2009; Decety & Cacioppo, 2011; Eberhardt, 2005; Ito & Bartholow, 2009; Kubota, et al., 2012; Todorov, et al., 2011; Van Bavel & Cunningham, 2011), only a few experiments have examined the functional neuroanatomy of perceiving and thinking about groups *qua* groups. However, this research gap is growing smaller. For example, recent reports have started to shed light on which brain regions represent stereotypes and other forms of semantic knowledge about social groups (Contreras, et al., 2012; Quadflieg, et al., 2009). Given the importance of groups to our everyday life as social animals, this burgeoning interest in the neural basis of intergroup cognition promises to become an important program of research in cognitive neuroscience.

STUDY 3

MULTIVOXEL PATTERNS IN FUSIFORM FACE AREA DIFFERENTIATE FACES BY SEX AND RACE

Contreras, J. M., Banaji, M. R., & Mitchell, J. P. (2013). Multivoxel patterns in fusiform face area differentiate faces by sex and race. *Under review at PLOS ONE*.

INTRODUCTION

One of the seminal breakthroughs in cognitive neuroscience was the discovery of a region of fusiform gyrus that responds preferentially to human faces, dubbed fusiform face area (FFA; Kanwisher, McDermott, & Chun, 1997; McCarthy, Puce, Gore, & Allison, 1997). FFA is thought to extract the physical information that distinguishes the faces of different people; that is, to represent face identity (for review, see Kanwisher & Yovel, 2006). Familiar faces elicit more neural activity in FFA than unrecognized faces (Grill-Spector, Knouf, & Kanwisher, 2004), and lesions to FFA impair face recognition (Barton, Press, Keenan, & O'Connor, 2002). Moreover, experiments using neural adaptation—in which repeated presentation of a stimulus property decreases neural activity in brain regions that represent the property (Grill-Spector & Malach, 2001)—suggest that FFA is more sensitive to changes in face identity than to physical changes unrelated to face identity (Andrews & Ewbank, 2004; Davies-Thompson, Newling, & Andrews, 2012; Rotshtein, Henson, Treves, Driver, & Dolan, 2005; cf. Xu, Yue, Lescroart, Biederman, & Kim, 2009).

But it is impossible to identify people by their faces without accurately categorizing their sex and race. The sex and race of a face determine how its identity is represented, inextricably linking face identity to these two social categories (for review, see Rhodes & Jaquet, 2011). Indeed, face morphology shows pronounced sexual dimorphism and racial differences (Farkas, Katic, & Forrest, 2005; Ferrario, Sforza, Pizzini, Vogel, & Miani, 1993). Recently, a set of studies have used multivariate pattern analysis (MVPA) to investigate whether fusiform gyrus represents the sex and race of faces. Univariate data analyses average the responses of multiple voxels. This spatial averaging reduces the information content of the data, which can exist at the level of the individual responses of multiple voxels, or *multivoxel patterns* (Kriegeskorte, et al., 2006). In contrast, MVPA interrogates these patterns to reveal the representations that a brain region contains (for review, see Weil & Rees, 2010). For example, a brain region in which faces of men and women elicit distinct multivoxel patterns but faces of the same sex yield similar patterns may represent sex.

Two studies have suggested that fusiform gyrus represents the sex and race of faces. In one study, participants in a functional magnetic resonance imaging (fMRI) scanner viewed faces of famous and unfamiliar men and women (Kaul, Rees, & Ishai, 2011). Pattern classifiers decoded the sex of the faces from fusiform gyrus. In another study, participants were scanned while viewing faces of unfamiliar Black and White individuals (Ratner, Kaul, & Van Bavel, 2012). Pattern classifiers decoded the race of the faces from fusiform gyrus. However, the sex finding has not been tested in FFA and the race finding has not been replicated reliably in FFA. Multivoxel patterns in FFA from participants who viewed the faces of Black and White individuals differentiated faces by race only for participants who showed high anti-Black bias (Brosch, Bar-David, & Phelps, 2012). A different study in which participants viewed photographs of Asian and White faces found that multivoxel patterns in FFA cannot distinguish faces by race (Natu, Raboy, & O'Toole, 2011). Therefore, these studies suggest that fusiform gyrus may represent sex and race. However, evidence on whether FFA represents race is mixed (one negative result and one qualified positive result) and no study of which we are aware has examined whether FFA represents sex.

Additionally, the studies that decoded social categories from fusiform gyrus (Brosch, et al., 2012; Kaul, et al., 2011; Ratner, et al., 2012) have an important limitation. They did not equate physical differences between photographs of social categories that were unrelated to their facial structure, such as luminance and contrast as well as high-level differences like hair length.

Consequently, the distinct patterns associated with social categories may not have reflected face differences. Consistent with this concern, the pattern classifiers in these studies decoded the social categories of faces in early visual cortex, which is not face-selective.

The present experiment continues the study of race representations in FFA and begins the study of sex representations in this face-selective brain region by scanning participants while they categorized faces of unfamiliar Black men, Black women, White men, and White women by sex and race. The goal of the present experiment is to determine if, despite the significant variability in the appearance of the people in the photographs, distinct pattern of voxels represent female and male faces as well as Black and White faces, suggesting that FFA includes representations of such social category information. We avoid the important limitation of insufficiently-controlled stimuli in two ways. First, we used photographs that are uniform in appearance and emotional expression, cropping face-irrelevant features (e.g., hairstyle) and background. Also, we controlled for low-level visual differences by equalizing luminance and contrast across social categories. Second, our stimuli orthogonalize sex and race so that if FFA differentiates faces by sex *and* race, this is unlikely to be caused by photograph differences unrelated to facial structure.

METHOD

Participants

Seventeen college students and community members from Cambridge, MA, participated in this study (9 female; age range 18-34, *M* = 22.18). All participants were right-handed, had no history of neurological problems, and described themselves as White. Participants provided their written informed consent in a manner approved for this study by the Committee on the Use of Human

Subjects in Research at Harvard University.

Stimuli and behavioral procedure

In a *categorization task*, participants viewed 192 photographs of unfamiliar Black men, Black women, White men, and White women (48 photographs in each condition). Because previous research is limited by insufficient stimuli control, the present stimuli were meticulously standardized to rule out alternative interpretations of any results. Photographs were collected from a variety of different online databases and depicted young adults facing forward with mouths closed, neutral expression, and eye gaze directed at the camera. The photographs were grayscaled and cropped to squares, their background was removed, and the luminance and contrast of the faces were equalized across conditions using in-house MATLAB code (MathWorks, Natick, MA). For example, the grayscaled images of Black and White faces differed in luminance ($M_{\text{Blacks}} = 106.67$, $M_{\text{Whites}} = 144.52$), $t(95) = 8.11$, $p < 10^{-12}$, but preprocessing removed this difference ($M_{\text{Blacks}} = 130$, $M_{\text{Whites}} = 130$).

In each scanning run, participants categorized the faces either by sex (man, woman) or by race (Black, White) using the index and middle fingers of their right hand, which rested on a button box. Each run was pseudorandomly assigned a categorization dimension (sex, race). Before each run, participants were instructed as to which categorization dimension (sex or race) to use and which button would correspond to each social category. Then, participants completed 10 practice trials on a set of 10 faces not used in the categorization task. Across runs, we counterbalanced the button assignments in such a way that each social category was assigned to each finger an equal number of times and each photograph was categorized once with the index finger and once with the middle finger.

Each trial lasted 2000 ms. For the first 500 ms, a photograph was shown in the center of the screen. For the remaining 1500 ms of each trial, the photograph was replaced with a white fixation crosshair, which encouraged participants to attend to the photographs closely. Photographs were segregated into 8 runs, each of which consisted of 48 photographs (12 in each of the four social categories, e.g., Black men). To optimize estimation of the event-related fMRI response, trials were intermixed in a pseudorandom order and separated by a variable stimulus interval (0-10 s) during which participants passively viewed a white fixation crosshair in the center of the screen (Dale, 1999).

After the categorization task, participants completed two runs of a canonical *face localizer* used to identify cortical regions responsive to faces (Kanwisher, et al., 1997). In each run, participants viewed photographs of human faces, human bodies, scenes, household objects, and scrambled versions of the household objects. Each photograph appeared for 1 s and was followed by a blank screen for 333 ms. Each category was blocked together to yield 10 blocks of 11 photographs each, 2 blocks per category. One photograph in each block was presented twice in a row, and participants were instructed to press a button when they detected this repetition. The blocks were separated by a stimulus interval that lasted 12 s and were presented in a pseudorandom order, such that participants could not anticipate the category of the upcoming block. During the task, participants fixated on a small, black circle that appeared in the center of the screen throughout the entire experiment (including the presentation of the photographs).

Functional imaging procedure

Imaging data were acquired on a 3.0 Tesla Siemens Tim Trio scanner (Siemens, Erlangen, Germany) with a standard head coil at the Center for Brain Science at Harvard University. Functional runs used a gradient-echo, echo-planar pulse sequence (TR = 3000 ms; TE = 28 ms; flip angle = 85°; field of view = 216 x 216 mm; matrix = 72 x 72; in-plane resolution = 2.5 x 2.5 mm; slice thickness = 2.5 mm). Forty-five interleaved axial slices parallel to the AC-PC line were obtained to cover most of the cerebrum; portions of superior parietal lobe were not covered. The categorization task consisted of 8 runs of 43 volume acquisitions each and the face localizer consisted of 2 runs of 98 volume acquisitions each. Each of the functional runs was preceded by 8 s of gradient and radio frequency pulses that allowed the scanner to reach steady-state magnetization. After the functional runs in the experiment, a high-resolution T1-weighted structural scan (MEMPRAGE) was conducted.

Functional imaging data analysis

Univariate analyses. FMRI data were preprocessed and analyzed using Statistical Parametric Mapping 8 (SPM8; Wellcome Department of Cognitive Neurology, London, United Kingdom) and in-house MATLAB code (MathWorks, Natick, MA) written by Dylan Wagner (Dartmouth College, Hanover, NH). To correct for head movement, a rigid-body transformation realigned images within each run and across all runs using the first functional image as a reference. Realigned images were unwarped to reduce any additional distortions caused by head movement. Unwarped data were normalized into a stereotaxic space (2-mm isotropic voxels) based on the SPM8 EPI template that conforms to the ICBM 152 brain template space and approximates the Talairach and Tournoux atlas space. Normalized images were spatially smoothed using a Gaussian kernel (8-mm full-width-at-half-maximum) to maximize signal-to-noise ratio and reduce the impact of individual differences in functional neuroanatomy. Finally, individual runs were analyzed on a participant-by-participant basis to find outlier volumes with Artifact Detection Toolbox (ART; McGovern Institute for Brain Research, Cambridge, MA). Outliers were defined as volumes in which participant head movement exceeded 0.5 mm or 1° and volumes in which overall signal were more than three standard deviations outside the mean global signal for the entire run.

For each participant, a general linear model (GLM) was constructed to include task effects and nuisance regressors (run mean, linear trend to account for signal drift over time, six movement parameters computed during realignment, and, if any, outlier scans identified by ART and trials in which participants did not provide a response). To compute unweighted (β) and weighted (t) parameter estimates for each condition at each voxel, the GLM was convolved with a canonical hemodynamic response function (HRF). The GLM of the categorization task was also convolved with the temporal and spatial derivatives of the HRF, which explain a significant portion of BOLD variability above and beyond the canonical model in event-related designs (Henson, et al., 2001). Trials were modeled as events of durations equal to their respective reaction times to account for differences in response times (RTs) across conditions (Grinband, et al., 2008).

Comparisons of interest were implemented as linear contrasts. In the categorization task, linear contrasts identified significant voxels with a voxel-wise statistical criterion of $p < .005$. Regions-of-interest (ROIs) were required to exceed 75 voxels in extent, establishing an experiment-wide statistical threshold of $p < .05$, corrected for multiple comparisons, on the basis of Monte Carlo simulations (Slotnick, et al., 2003). In the face localizer, ROIs were identified for each participant with a voxel-wise statistical criterion of, at most, $p < .05$ (median $p = .005$). Additional statistical comparisons between conditions were conducted in MATLAB using

ANOVA on the parameter estimates associated with each trial type.

Multivariate analyses. Preprocessing and GLM estimation were identical to those for the univariate analysis of the face categorization task, except that normalized images were spatially smoothed using a smaller Gaussian kernel (5-mm full-width-at-half-maximum).

Trials were conditionalized by sex (men, women), race (Black, White) and run type (odd, even) to yield eight conditions (e.g., *Black men-even*). Linear contrasts compared each condition to baseline. Following Misaki, Kim, Bandettini, and Kriegeskorte (2010), these parameter estimates were used for the rest of the analysis to reduce the influence of noisy voxels. The parameter estimates were extracted from each of the ROIs defined by the face localizer and correlated in three ways: same-sex correlations (*Black men-odd with White men-even, Black men-even with White men-odd, Black women-odd with White women-even, Black women-even with White women-odd*), same-race correlations (*Black men-odd with Black women-even, Black men-even with Black women-odd, White men-odd with White women-even, White men-even with White women-odd*), and different-category correlations (*Black men-odd with White women-even, White men-odd with Black women-even, Black women-odd with White men-even, White women-odd with Black men-even*).

Correlations were Fisher-transformed to *z*-values and averaged to yield one same-sex correlation, one same-race correlation, and one different-category correlation. Then, the different-category correlation was subtracted from each of the other average correlations to yield two correlation differences. Finally, one-tailed, one-sample *t*-tests determined if these correlation differences were reliably greater than zero across participants.

RESULTS

Behavioral data

Table 3.1 displays means and standard deviations of responses and RTs. Participants categorized faces more accurately and more quickly by sex (M_{accuracy} = 0.98, M_{RT} = 670 ms) than race (M_{accuracy} = 0.95, M_{RT} = 712 ms), *ts*(16) > 5.65, *ps* < 10^{-5}, *Cohen's d*s > 1.41. Participants categorized men (M_{accuracy} = 0.97, M_{RT} = 684 ms) more accurately and more quickly than women (M_{accuracy} = 0.96, M_{RT} = 699 ms), *ts*(16) > 2.25, *ps* < .04, *d*s > 0.56. Although participants were no more accurate to categorize Black (M_{accuracy} = 0.96) than White faces (M_{accuracy} = 0.96), *p* = .15, they were faster to categorize Black (M_{RT} = 683 ms) than White faces (M_{RT} = 699 ms), *t*(16) = 3.05, *p* < .01, *d* = 0.76. The sex and race of photographs did not interact in participants' accuracy and RT, whether collapsing across sex and race runs, within sex runs, or within race runs, all *p*s > .22. Moreover, the 3-way interaction of photograph sex, photograph race, and run (sex, race) was not statistically reliable for accuracy and RT, all *p*s > .28.

Functional imaging data

Univariate analyses. The face localizer was used to identify FFA and control brain regions independently (Table 3.2). Replicating previous research (Kanwisher, et al., 1997; McCarthy, et al., 1997), the contrast of *faces* > [*bodies* + *scenes* + *objects* + *scrambled objects*] identified a bilateral region of fusiform gyrus that corresponds to FFA. As face-selective control regions, this contrast also identified a bilateral region of inferior occipital gyrus that corresponds to occipital face area (OFA) (Gauthier et al., 2000), and a bilateral region of superior temporal sulcus (STS) (Puce, Allison, Bentin, Gore, & McCarthy, 1998). As control regions that are category-selective but not face-selective, the contrast of *scenes* > *objects* identified a bilateral

Table 3.1 Participants' responses and response latencies from the categorization task.

	Accuracies		*Response Latencies*	
	Sex	*Race*	*Sex*	*Race*
White men	0.95^{ac} (0.04)	0.98^{b} (0.02)	706^{acd} (65)	679^{b} (64)
White women	0.94^{c} (0.05)	0.97^{ad} (0.03)	722^{cd} (86)	692^{ab} (78)
Black men	0.96^{acd} (0.04)	0.98^{b} (0.03)	700^{ac} (60)	650^{e} (65)
Black women	0.94^{c} (0.05)	0.98^{bd} (0.02)	722^{d} (67)	661^{f} (55)

Note: Means and, in parentheses, standard deviations. Accuracies are displayed in proportions of correct categorizations. Response times are displayed in milliseconds. For each dependent variable, means sharing a superscript do not differ significantly at $p < .05$, as computed in paired-samples *t*-tests.

Table 3.2 Brain regions identified in whole-brain, random-effects contrasts in the categorization task, *p* < .05, corrected for multiple comparisons.

Region	*x*	*y*	*z*	*Participants*
		Faces > [Bodies + Scenes + Objects + Scrambled Objects]		
Fusiform gyrus (FFA)	38.8	-44.3	-18.5	16
	-37.1	-47.6	-17.3	16
Inferior occipital gyrus (OFA)	33.3	-76.7	-8.9	14
	-33.1	-77.0	-6.55	11
Superior temporal sulcus (STS)	49.8	-43.4	13.9	16
	-49.8	-52.8	21.3	9
		Scenes > Objects		
Parahippocampal gyrus (PPA)	23.4	-39.5	-7.4	16
	-24.1	-42.9	-4.8	16
		Objects > Scrambled Objects		
Lateral occipital cortex (LOC)	40.5	-66.3	-5.0	8
	-42.0	-63.7	-6.7	10

Note: From left to right, columns list the names of regions obtained from whole-brain, random-effects contrasts, the mean stereotaxic Montreal Neurological Institute coordinates of their peak voxels across participants, and the number of participants (*N* = 17) in whom these brain regions were identified at *p* < .05, corrected for multiple comparisons. FFA = fusiform face area, OFA = occipital face area, STS = superior temporal sulcus, PPA = parahippocampal place area, LOC = lateral occipital complex.

region of parahippocampal gyrus that corresponds to parahippocampal place area (PPA; Epstein & Kanwisher, 1998). Additionally, the contrast of *objects* > *scrambled objects* identified a bilateral region of lateral occipital cortex that corresponds to lateral occipital complex (LOC; Malach et al., 1995).

For completeness, univariate analyses of the categorization task examined potential differences between photographs as a function of their sex and race. For these analyses, trials were conditionalized by sex (men, women) and race (Black, White; Table 3.3).

Multivariate analyses. First, we examined whether FFA maintains distinct representations of female and male faces; that is, whether multivoxel patterns in FFA show higher correlations between photographs of individuals of the same sex than between photographs of men and women (Figure 3.1). Consistent with the hypothesis that FFA distinguishes faces by sex, pattern correlations in FFA were higher between photographs of the same sex than between photographs of men and women (right FFA, $t(15) = 3.03$, $p < .005$; left FFA, $t(15) = 2.73$, $p < .008$). The correlation differences of right and left FFA were equivalent, $t(14) = 0.69$, $p = 0.50$, suggesting that both regions distinguished faces by sex to a similar degree.

Second, we examined whether FFA maintains distinct representations of Black and White faces; that is, whether multivoxel patterns in FFA show higher correlations between photographs of individuals of the same race than between photographs of Black and White individuals (Figure 3.1). Consistent with the hypothesis that FFA distinguishes faces by race, pattern correlations in FFA were higher between photographs of the same race than between photographs of Black and White faces (right FFA, $t(15) = 1.72$, $p = .05$; left FFA, $t(15) = 2.21$, $p < .02$). The correlation differences of right and left FFA were equivalent, $t(14) = 1.01$, $p = 0.33$, suggesting that both

Table 3.3 Brain regions identified in whole-brain, random-effects contrasts in the face localizer task, *p* < .05, corrected for multiple comparisons, sorted in descending order by the *t*-statistic of their peak voxel (*t*).

Region	*x*	*y*	*z*	*k*	*t*
			Men > Women		
			No brain regions identified.		
			Women > Men		
Cerebellum	0	-61	-16	204	5.18
Inferior frontal gyrus	-28	15	-20	231	4.71
Superior frontal gyrus	20	61	-6	89	4.50
Cingulate gyrus	4	-29	34	75	3.99
			White > Black		
Middle frontal gyrus	-16	33	-8	437	7.73
	14	35	-12	162	6.06
Cerebellum	-12	-57	-32	82	5.08
Cingulate gyrus	-20	-31	44	112	4.83
Precuneus	-16	-45	22	105	4.12
			Black > White		
White matter	-18	-81	2	126	5.36
Supramarginal gyrus	48	-53	34	142	4.60

Note: From left to right, columns list the names of regions obtained from whole-brain, random-effects contrasts, the stereotaxic Montreal Neurological Institute coordinates of their peak voxels, their size in number of voxels (*k*), and the *t*-statistic of their peak voxel (*t*).

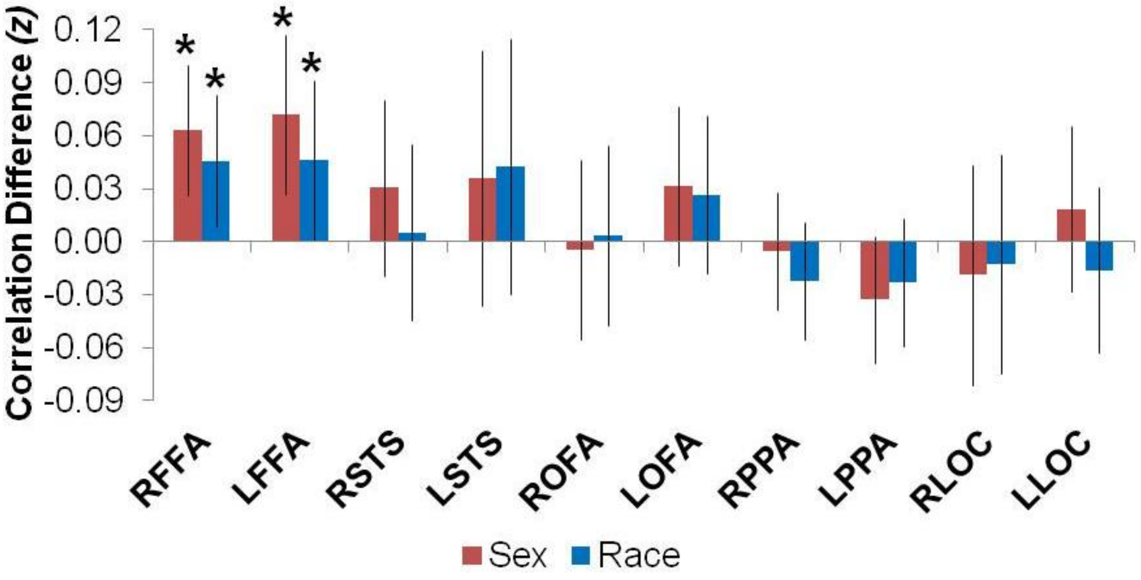

Figure 3.1 Bar graphs display mean correlation differences expressed in *z*-scores (*same-sex* > *different-category* in red, *same-race* > *different-category* in blue). An asterisk denotes a correlation difference that is reliably greater than zero across participants, *p* < .05. Error bars represent 95% confidence intervals in within-subject comparisons (Masson & Loftus, 2003). R and L as the first letters of a region-of-interest's (ROI) acronym denote the brain hemisphere in which the ROI is localized. FFA = fusiform face area, OFA = occipital face area, STS = superior temporal sulcus, PPA = parahippocampal place area, LOC = lateral occipital complex.

regions distinguished faces by race to a similar degree.

We speculated that FFA might be the only face-selective brain region to represent the sex and race of faces because it is the face-selective region that is most sensitive to face identity (Kanwisher & Yovel, 2006). To test this hypothesis, we repeated the MVPA with patterns extracted from other brain regions defined by the face localizer, which included ones previously implicated in face processing like OFA and STS (Kanwisher & Yovel, 2006) (Figure 3.1). Neither right nor left OFA or STS distinguished faces by social category reliably, *p*s > .13. This suggests that FFA is alone among face-selective brain regions in decoding the sex and race of faces. Because face information may exist in category-selective cortex outside of FFA (Haxby, et al., 2001; Op de Beeck, Brants, Baeck, & Wagemans, 2010), we repeated the pattern similarity analyses with patterns extracted from place-selective PPA and object-selective LOC (Figure 3.1). Neither right nor left PPA or LOC distinguished faces by social category reliably, *p*s > .26. This suggests that other category-selective brain regions lack sex and race information about faces.

However, FFA may differentiate photographs not by facial properties that vary between social categories, but by lower-level physical differences between the photographs. Many of these low-level physical differences were removed by careful photograph selection and intensive preprocessing (see *Method: Stimuli and behavioral procedure*), but we wanted to test this alternative hypothesis empirically. Therefore, we analyzed multivoxel patterns from early visual cortex, which processes lower-level visual features. To do so, we used the stereotaxic coordinates of the center of mass of the right ([*x y z*] = 25, -82, -15) and left ([*x y z*] = -29, -80, -18) foveal confluence of brain areas V1, V2, and V3, which represents the central portion of the visual field, as functionally-defined by Dougherty *et al.* (2003) using retinotopic mapping (Engel et al., 1994). We extracted patterns from 8-mm spheres centered on these stereotaxic coordinates and repeated the pattern similarity analyses with these patterns. Neither the right nor the left foveal confluence distinguished faces by social category reliably, *p*s > .66. This suggests that low-level visual differences between the photographs do not cause multivoxel patterns in FFA to differentiate faces by sex and race.

Finally, we tested for effects of categorization dimension. To do so, trials were conditionalized by sex (men, women), race (Black, White), categorization dimension (sex, race), and run type (odd, even) to yield 16 conditions (e.g., *Black men categorized by sex-even*). Then, the same correlation differences as before (*same-sex > different-category, same-race > different-category*) were calculated separately for each categorization dimension (e.g., *same-sex categorized by sex > different-category categorized by sex*). None of these correlation differences were reliably larger than zero, *p*s > .16. The discrepancy between these results and the positive results of the analysis in which trials were not conditionalized by categorization dimension are most likely caused by differences in statistical power. The analysis that involves conditionalizing by categorization dimension has half as many trials per condition as the other analysis, endowing it with an inferior ability to detect small differences between multivoxel patterns across conditions.

DISCUSSION

Previous studies suggested that fusiform gyrus represents the sex and race of faces (Kaul, et al., 2011; Ratner, et al., 2012), although whether FFA in particular represents this information was unclear (Brosch, et al., 2012; Natu, et al., 2011). In the present experiment, we observed that multivoxel patterns in bilateral FFA distinguished faces by sex and race. Participants variably categorized photographs of unfamiliar Black men, Black women, White men, and White women by sex and race. Despite the significant variability in the appearance of the people in the photographs, a distinct pattern of voxels distinguished between female and male faces and between Black and White faces, suggesting that bilateral FFA includes representations of such social category information.

These social category representations may be components of face identity representations, which are thought to exist in FFA (Kanwisher & Yovel, 2006). Because face identity is inextricably linked to social categories like age, sex, and race (for review, see Rhodes & Jaquet, 2011), it seems reasonable that FFA might represent face identity as well as the social categories of faces. FFA could be the neuroanatomical locus in which social categories that are relevant to face identity (i.e., age, race, and sex) are integrated to form holistic representations of individual faces. This hypothesis is consistent with behavioral research that suggests that the human brain codes face identity with reference to social categories (Rhodes & Leopold, 2011).

Analyses of multivoxel patterns from other brain regions suggest that representations of the sex and race of faces may be unique to FFA. Patterns extracted from other face-selective brain regions (OFA and STS), other category-selective brain regions (PPA and LOC), and early visual cortex (foveal confluence of V1, V2, and V3) did not differentiate faces by sex or race. The null results from patterns in early visual cortex suggest that the careful selection and intensive preprocessing of the stimuli removed low-level physical differences unrelated to the sex and race of the stimuli that might have existed in the original photographs. These null results are especially important in this experiment because previous studies that decoded the sex or race of faces from fusiform gyrus also decoded sex and race from early visual cortex (Brosch, et al., 2012; Kaul, et al., 2011; Ratner, et al., 2012).

FFA is thought to process perceptual rather than semantic aspects of person perception

(Kanwisher & Yovel, 2006). For this reason, the sex and race information that FFA represents is unlikely to be semantic; that is, FFA may “tell” faces apart by sex and race without “knowing” what these differences mean. Nonetheless, FFA may play a critical role in social categorization. Indeed, multivoxel patterns in FFA may be not only stimulus-specific (male faces vs. female faces) but also task-specific (sex vs. race categorization; Chiu, Esterman, Han, Rosen, & Yantis, 2011). This suggests that neural activity in FFA can be modulated by the type of social categorization that a person performs. One of the most fruitful future directions for research on sex and race representations in FFA may be to investigate how this information guides semantic retrieval about social categories in more anterior regions of temporal lobe, which have been consistently implicated in semantics about people generally (for review, see Wong & Gallate, 2012) and in stereotypes specifically (Contreras, et al., 2012). Evidence exists to suggest that stereotyping can modulate neural activity in FFA (Quadflieg et al., 2011), but how representations in FFA might inform higher-order social processes like stereotyping is unknown.

In sum, the present experiment suggests that FFA distinguishes faces by social categories like sex and race. In this way, the current research contributes to our emerging understanding of how the human brain perceives individuals from different social categories.

DISCUSSION

Though social groups shape almost every respect of our lives, cognitive neuroscientists have not conducted many investigations to study how the human brain solves the computational challenges that allow human beings to navigate a social world fragmented along the lines of group membership. The studies presented in this dissertation explore three domains of intergroup cognition that have not received much, if any, attention from cognitive neuroscientists: Semantic knowledge, theory of mind about groups, and social categorization.

Study 1 finds that brain regions previously identified in processing semantic information are more robustly engaged by nonsocial semantics than stereotypes. In contrast, stereotypes elicit greater activity in brain regions implicated in social cognition. These results suggest that stereotypes should be considered distinct from other forms of semantic knowledge. Study 2 shows that this same set of brain regions is more robustly responsive to inferences about the mental states rather than physical aspects of groups of people. However, this study also finds that multivoxel patterns in these brain regions differentiate groups from individual members. These findings suggest that perceivers mentalize about groups in a manner qualitatively similar to mentalizing about individual people, but that the brain nevertheless maintains important distinctions between the representations of such entities. Finally, Study 3 shows that patterns of voxel-based responses differentiate between the faces of Blacks and Whites and between the faces of men and women. These results suggest that a face-selective brain region in the visual system have distinct representations of the sex and race of faces.

In addition to addressing important domains of intergroup cognition not previously explored in previous cognitive neuroscience research, these studies contribute to our understanding of how the brain enables intergroup cognition in two other ways. First, Studies 1 and 2 examine the functional neuroanatomy of perceiving and thinking about groups *qua* groups rather than examining how we perceive other individuals as group members. As the study of the neural basis of social psychology progresses, it will need to devote greater attention to the perception of groups as units of analysis distinct from the individuals that comprise them. Second, Studies 2 and 3 attempted to identify brain regions that contain distinct representations of groups and individuals and distinct representations of different social groups. Again, the study of the neural basis of social psychology will require probing not only which brain regions are involved in thinking about groups, but understanding what group information these brain regions represent and how this information is used.

Future research in the cognitive neuroscience of social groups should address the following questions left unanswered by the present studies. First, what makes semantic knowledge about individuals and groups similar? The retrieval of semantic knowledge about individuals recruits similar brain regions to those identified in Study 1 (Mitchell, et al., 2002; Zahn et al., 2007), suggesting that semantic knowledge about individual people and the groups that they form share important qualities. Indeed, the representational spaces that support semantic knowledge about individuals and groups share organizing principles. The contents of stereotypes and interpersonal perceptions are arranged in a two-dimensional space of warmth and competence (for review, see Fiske, Cuddy, & Glick, 2007). Using multivariate pattern analysis, a future experiment could determine whether the brain regions identified in Study 1 have a common neural code of warmth and competence that underlie multivoxel patterns that distinguish groups from each other and individuals from each other along dimensions like warmth and competence.

Second, collections of people vary in the degree to which they are considered cohesive groups (Lickel et al., 2000). Moreover, group cohesiveness influences the degree to which we attribute a mind to a group and the kind of mind that we attribute to a group (Knobe & Prinz, 2008; Waytz & Young, 2012). Therefore, Study 2's finding that theory of mind about individuals and groups has a common neural basis may be moderated by the cohesiveness of social groups. Inferences about the mental states of cohesive groups may recruit brain regions implicated in theory of mind to a greater degree than inferences about groups that are not cohesive entities. Alternatively, it may be the case that cohesive and non-cohesive groups can be distinguished in these brain regions not by the amount of activity that they elicit, but instead by the multivoxel patterns with which they are associated. Study 2 suggests that these brain regions differentiate groups from individuals, but they leave open the question of whether these brain regions can discriminate between different groups.

Finally, social categorization of sex and race involves not only face perception, but also body perception. Men and women differ from each other not only by facial structure, but also by body shape. Likewise, people from different races differ in their bodily appearance. Study 3 suggests that the visual system contains distinct representations of the sex and race of faces, but whether it also discriminates bodies by sex and race is unknown. Future research may investigate this question by examining whether extrastriate body area, a functionally-defined region of lateral occipitotemporal cortex that processes information about body form (for review, see Peelen & Downing, 2007), has multivoxel patterns that differentiate bodies by sex and race. Further work in this area could study how information about the sex and race of faces and bodies is integrated into a coherent representation of the social category to which a person belongs.

Together, the three studies that comprise this dissertation aim to increase our understanding of the functional neuroanatomy that underlies the human capacity to think about social groups. Given the importance of groups to our everyday life as social animals, the burgeoning interest of cognitive neuroscientists in the neural basis of intergroup cognition promises to become an important program of research in the study of the neural basis of human psychology.

REFERENCES

Allport, G. (1954). *The nature of prejudice*. Cambridge, MA: Addison-Wesley.

Ames, D. R. (2004a). Inside the mind-reader's tool kit: Projection and stereotyping in mental state inference. *Journal of Personality and Social Psychology, 87*(3), 340-353.

Ames, D. R. (2004b). Strategies for social inference: A similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of Personality and Social Psychology, 87*(5), 573-585.

Ames, D. R. (2005). Everyday solutions to the problem of other minds. In B. F. Malle & S. D. Hodges (Eds.), *Other minds: How human bridge the divide between self and others* (pp. 158-173). New York, NY: Guilford Publications.

Ames, D. R., & Mason, M. F. (2012). Mind perception. In S. T. Fiske & C. N. Macrae (Eds.), *The SAGE handbook of social cognition* (pp. 115-137). Thousand Oaks, CA: Sage.

Ames, D. R., Weber, E. U., & Zou, X. (2012). Mind-reading in strategic interaction: The impact of perceived similarity on projection and stereotyping. *Organizational Behavior and Human Decision Processes, 117*(1), 96-110.

Amodio, D. M., Harmon-Jones, E., & Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink response and self-report. *Journal of Personality and Social Psychology, 84*(4), 738-753.

Andrews, T. J., & Ewbank, M. P. (2004). Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *NeuroImage*, *23*(3), 905-913.

Aristotle. (1975). *Categories and de interpretatione* (J. L. Ackrill, Trans.). Oxford: Oxford University Press.

Baldo, J. V., & Shimamura, A. P. (1998). Letter and category fluency in patients with frontal lobe lesions. *Neuropsychology, 12*(2), 259-267.

Banaji, M. R., & Bhaskar, R. (1999). Implicit stereotypes and memory: The bounded rationality of social beliefs. In D. L. Schacter & E. Scarry (Eds.), *Memory, brain, and belief* (pp. 139-175). Cambridge, MA: Harvard University Press.

Barton, J. J., Press, D. Z., Keenan, J. P., & O'Connor, M. (2002). Lesions of the fusiform face area impair perception of facial configuration in prosopagnosia. *Neurology, 58*(1), 71-78.

Beer, J. S., Stallen, M., Lombardo, M. V., Gonsalkorale, K., Cunningham, W. A., & Sherman, J. W. (2008). The Quadruple Process model approach to examining the neural underpinnings of prejudice. *NeuroImage, 43*(4), 775-783.

Bookheimer, S. (2002). Functional MRI of language: new approaches to understanding the cortical organization of semantic processing. *Annual Review of Neuroscience, 25*, 151-188.

Bottom, W. P., & Paese, P. W. (1997). False consensus, stereotypic cues, and the perception of integrative potential in negotiation. *Journal of Applied Social Psychology, 27*(21), 1919-1940.

Brewer, M. B., & Harasty, A. S. (1996). Seeing groups as entities: The role of perceiver motivation. In R. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition* (Vol. 3, pp. 347-370). New York City: Guilford.

Brosch, T., Bar-David, E., & Phelps, E. A. (2012). Implicit race bias decreases the similarity of the neural representations of Black and White faces. *Psychological Science*.

Brunet, E., Sarfati, Y., Hardy-Baylé, M. C., & Decety, J. (2000). A PET investigation of the attribution of intentions with a nonverbal task. *NeuroImage, 11*(2), 157-166.

Cantor, N., & Mischel, W. (1979). Prototypicality and personality: Effects on free-recall and personality impressions. *Journal of Research in Personality, 13*(2), 187-205.

Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience, 10,* 1-34.

Castelli, F., Happe, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage, 12*(3), 314-325.

Centelles, L., Assaiante, C., Nazarian, B., Anton, J. L., & Schmitz, C. (2011). Recruitment of both the mirror and the mentalizing networks when observing social interactions depicted by point-lights: A neuroimaging study. *PLOS ONE, 6*(1), e15749.

Chiu, Y. C., Esterman, M., Han, Y., Rosen, H., & Yantis, S. (2011). Decoding task-based attentional modulation during face categorization. *Journal of Cognitive Neuroscience, 23*(5), 1198-1204.

Contreras, J. M., Banaji, M. R., & Mitchell, J. P. (2012). Dissociable neural correlates of stereotypes and other forms of semantic knowledge. *Social Cognitive & Affective Neuroscience, 7*(7), 764-770.

Cunningham, W. A., Raye, C. L., & Johnson, M. K. (2004). Implicit and explicit evaluation: FMRI correlates of valence, emotional intensity, and control in the processing of attitudes. *Journal of Cognitive Neuroscience, 16*(10), 1717-1729.

Cunningham, W. A., & Van Bavel, J. J. (2009). A neural analysis of intergroup perception and evaluation. In G. G. Berntson & J. T. Cacioppo (Eds.), *Handbook of neuroscience for the behavioral sciences* (pp. 975-984). New York City: John Wiley & Sons.

Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Human Brain Mapping, 8*(2-3), 109-114.

Davies-Thompson, J., Newling, K., & Andrews, T. J. (2012). Image-invariant responses in face-selective regions do not explain the perceptual advantage for familiar face recognition. *Cerebral Cortex*.

Decety, J., & Cacioppo, J. T. (2011). *The Oxford handbook of social neuroscience*. New York City: Oxford University Press.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dougherty, R. F., Koch, V. M., Brewer, A. A., Fischer, B., Modersitzki, J., & Wandell, B. A. (2003). Visual field representations and locations of visual areas V1/2/3 in human visual cortex. *Journal of Vision, 3*(10), 586-598.

Dovidio, J. F., & Gaertner, S. L. (2010). Intergroup bias. In D. T. Gilbert, S. T. Fiske & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 2, pp. 1084-1121). New York City: Wiley.

Eberhardt, J. L. (2005). Imaging race. *American Psychologist, 60*(2), 181-190.

Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E. J., & Shadlen, M. N. (1994). fMRI of human visual cortex. *Nature, 369*(6481), 525.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864-886.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598-601.

Farkas, L. G., Katic, M. J., & Forrest, C. R. (2005). International anthropometric study of facial morphology in various ethnic groups/races. *Journal of Craniofacial Surgery, 16*(4), 615-646.

Ferrario, V. F., Sforza, C., Pizzini, G., Vogel, G., & Miani, A. (1993). Sexual dimorphism in the human face assessed by euclidean distance matrix analysis. *Journal of Anatomy, 183*(3), 593-600.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77-83.

Fletcher, P., Happe, F., Frith, U., Baker, S., Dolan, R., Frackowiak, R., & Frith, C. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition, 57*(2), 109-128.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*(4), 531-534.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences, 7*(2), 77-83.

Gallagher, H. L., Happe, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: An fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia, 38*(1), 11-21.

Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage, 16*(3), 814-821.

Gauthier, I., Tarr, M. J., Moylan, J., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). The fusiform "face area" is part of a network that processes faces at the individual level. *Journal of Cognitive Neuroscience, 12*(3), 495-504.

Goel, V., Grafman, J., Sadato, N., & Hallett, M. (1995). Modeling other minds. *Neuroreport, 6*(13), 1741.

Golby, A. J., Gabrieli, J. D., Chiao, J. Y., & Eberhardt, J. L. (2001). Differential responses in the fusiform region to same-race and other-race faces. *Nature Neuroscience, 4*(8), 845-850.

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nature Neuroscience, 7*(5), 555-562.

Grill-Spector, K., & Malach, R. (2001). fMR-adaptation: A tool for studying the functional properties of human cortical neurons. *Acta Psychologica, 107*, 293-321.

Grinband, J., Wager, T. D., Lindquist, M., Ferrera, V. P., & Hirsch, J. (2008). Detection of time-varying signals in event-related fMRI designs. *NeuroImage, 43*(3), 509-520.

Hackman, J. R., & Katz, N. (2010). Group behavior and performance. In D. T. Gilbert, S. T. Fiske & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 2, pp. 1084-1121). New York City: Wiley.

Hamilton, D. (1981). *Cognitive processes in stereotyping and intergroup behavior*. Hillsdale, NJ: Erlbaum.

Hamilton, D. L., & Sherman, S. J. (1996). Perceiving persons and groups. *Psychological Review, 103*(2), 336-355.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425-2430.

Henson, R., Rugg, M. D., & Friston, K. J. (2001). The choice of basis functions in event-related fMRI. *NeuroImage, 13*(6), S149-S149.

Hill, R. A., & Dunbar, R. I. M. (2003). Social network size in humans. *Human Nature, 14*(1), 53-72.

Hodges, J. R., Patterson, K., Oxbury, S., & Funnell, E. (1992). Semantic dementia: Progressive fluent aphasia with temporal lobe atrophy. *Brain, 115*(6), 1783-1806.

Iacoboni, M., Lieberman, M. D., Knowlton, B. J., Molnar-Szakacs, I., Moritz, M., Throop, C. J., & Fiske, A. P. (2004). Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. *NeuroImage, 21*(3), 1167-1173.

Ikemoto, S., & Panksepp, J. (1999). The role of nucleus accumbens dopamine in motivated behavior: A unifying interpretation with special reference to reward-seeking. *Brain Research Reviews, 31*(1), 6-41.

Ito, T. A., & Bartholow, B. D. (2009). The neural correlates of race. *Trends in Cognitive Sciences, 13*(12), 524-531.

Jenkins, A. C., Macrae, C. N., & Mitchell, J. P. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proceedings of the National Academy of Sciences United States of America, 105*(11), 4507-4512.

Joiner, R., Gavin, J., Brosnan, M., Cromby, J., Gregory, H., Guiller, J., . . . Moon, A. (2012). Gender, internet experience, Internet identification, and internet anxiety: A ten-year followup. *Cyberpsychology, Behavior, and Social Networking, 15*(7), 370-372.

Jones, T. E. (2010). *What people believe when they say that people believe: Folk sociology and the nature of group intentions*. Lanham, MD: Lexington Books.

Joseph, J. E. (2001). Functional neuroimaging studies of category specificity in object recognition: A critical review and meta-analysis. *Cognitive, Affective, & Behavioral Neuroscience, 1*(2), 119-136.

Kandler, C., Bleidorn, W., & Riemann, R. (2012). Left or right? Sources of political orientation: The roles of genetic factors, cultural transmission, assortative mating, and personality. *Journal of Personality and Social Psychology, 102*(3), 633-645.

Kant, I. (1781/2003). *Critique of pure reason* (N. K. Smith, Trans.). New York City: Palgrave Macmillan.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302-4311.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences, 361*(1476), 2109-2128.

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).

Knobe, J., & Prinz, J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences, 7*(1), 67-83.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLOS ONE, 3*(7), e2597.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences United States of America, 103*(10), 3863-3868.

Kubota, J. T., Banaji, M. R., & Phelps, E. A. (2012). The neuroscience of race. *Nature Reviews Neuroscience*.

Lickel, B., Hamilton, D. L., Wieczorkowska, G., Lewis, A., Sherman, S. J., & Uhles, A. N. (2000). Varieties of groups and the perception of group entitativity. *Journal of Personality and Social Psychology*, *78*(2), 223-246.

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*(6), 720-722.

Lingle, J. H., Altom, M. W., & Medin, D. L. (1984). Of cabbages and kings: Assessing the extendability of natural object concept models to social things. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 1, pp. 71-118). Hillsdale, NJ: Lawrence Erlbaum.

Ly, M., Haynes, M. R., Barter, J. W., Weinberger, D. R., & Zink, C. F. (2011). Subjective socioeconomic status predicts human ventral striatal responses to social status information. *Current Biology, 21*(9), 794-797.

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., . . . Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences United States of America, 92*(18), 8135-8139.

Martin, A. (2001). Functional neuroimaging of semantic memory. In R. Cabeza & A. Kingstone (Eds.), *Handbook of functional neuroimaging of cognition* (pp. 153-186). Cambridge, MA: MIT Press.

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: Structure and processes. *Current Opinion in Neurobiology, 11*(2), 194-201.

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57*(3), 203-220.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences United States of America, 98*(20), 11832-11835.

McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience, 9*, 605-610.

Medin, D. L., & Smith, E. E. (1984). Concepts and concept formation. *Annual Review of Psychology, 35*, 113-138.

Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage, 53*(1), 103-118.

Mitchell, J. P. (2009a). Inferences about mental states. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1521), 1309-1316.

Mitchell, J. P. (2009b). Social psychology as a natural kind. *Trends in Cognitive Sciences, 13*(6), 246-251.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). General and specific contributions of the medial prefrontal cortex to knowledge about mental states. *NeuroImage, 28*(4), 757-762.

Mitchell, J. P., Cloutier, J., Banaji, M. R., & Macrae, C. N. (2006). Medial prefrontal dissociations during processing of trait diagnostic and nondiagnostic person information. *Social Cognitive and Affective Neuroscience, 1*(1), 49-55.

Mitchell, J. P., Heatherton, T. F., & Macrae, C. N. (2002). Distinct neural systems subserve person and object knowledge. *Proceedings of the National Academy of Sciences United States of America, 99*(23), 15238-15243.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2004). Encoding specific effects of social cognition on the neural correlates of subsequent memory. *Journal of Neuroscience, 24*(21), 4912-4917.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2005). Forming impressions of people versus inanimate objects: Social-cognitive processing in the medial prefrontal cortex. *NeuroImage, 26*, 251-257.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50*(4), 655-663.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Natu, V., Raboy, D., & O'Toole, A. J. (2011). Neural correlates of own- and other-race face perception: Spatial and temporal response differences. *NeuroImage, 54*(3), 2547-2555.

Norris, C. J., Chen, E. E., Zhu, D. C., Small, S. L., & Cacioppo, J. T. (2004). The interaction of social and emotional processes in the brain. *Journal of Cognitive Neuroscience, 16*(10), 1818-1829.

Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The enigmatic temporal pole: A review of findings on social and emotional processing. *Brain, 130*(7), 1718-1731.

Op de Beeck, H. P., Brants, M., Baeck, A., & Wagemans, J. (2010). Distributed subordinate specificity for bodies, faces, and buildings in human ventral visual cortex. *NeuroImage, 49*(4), 3414-3425.

Ostrom, V. (1984). The meaning of value terms. *American Behavioral Scientist, 28*(2), 249-262.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Review Neuroscience, 8*(12), 976-987.

Peelen, M. V., & Downing, P. E. (2007). The neural basis of visual body perception. *Nature Reviews Neuroscience, 8*(8), 636-648.

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. *Journal of Cognitive Neuroscience, 12*(5), 729-738.

Platek, S. M., & Krill, A. L. (2009). Self-face resemblance attenuates other-race face effect in the amygdala. *Brain Research, 1284,* 156-160.

Plous, S. (1993). The nuclear arms race: Prisoner's dilemma or perceptual dilemma? *Journal of Peace Research, 30*(2), 163-179.

Poldrack, R. A., Wagner, A. D., Prull, M. W., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. E. (1999). Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. *NeuroImage, 10,* 15-35.

Povinelli, D. J., & Preuss, T. M. (1995). Theory of mind: Evolutionary history of a cognitive specialization. *Trends in Neurosciences, 18*(9), 418-424.

Premack, D. G., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1,* 515-526.

Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *Journal of Neuroscience, 18*(6), 2188-2199.

Quadflieg, S., Flannigan, N., Waiter, G. D., Rossion, B., Wig, G. S., Turk, D. J., & Macrae, C. N. (2011). Stereotype-based modulation of person perception. *NeuroImage, 57*(2), 549-557.

Quadflieg, S., Turk, D. J., Waiter, G. D., Mitchell, J. P., Jenkins, A. C., & Macrae, C. N. (2009). Exploring the neural correlates of social stereotyping. *Journal of Cognitive Neuroscience, 21*(8), 1560-1570.

Ratner, K. G., Kaul, C., & Van Bavel, J. J. (2012). Is race erased? Decoding race from patterns of neural activity when skin color is not diagnostic of group boundaries. *Social Cognitive & Affective Neuroscience*.

Rhodes, G., & Jaquet, E. (2011). Aftereffects reveal that adaptive face-coding mechanisms are selective for race and sex. In R. A. A. Jr., N. Ambady, K. Nakayama & S. Shimojo (Eds.), *The science of social vision* (pp. 347-362). New York City: Oxford University Press.

Rhodes, G., & Leopold, D. A. (2011). Adaptive norm-based coding of face identity. In A. J. Calder, G. Rhodes, M. H. Johnson & J. V. Haxby (Eds.), *The Oxford handbook of face perception* (pp. 263-286). New York City: Oxford University Press.

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. *Nature Neuroscience, 6*(12), 1323-1328.

Ronquillo, J., Denson, T. F., Lickel, B., Lu, Z. L., Nandy, A., & Maddox, K. B. (2007). The effects of skin tone on race-related amygdala activity: An fMRI investigation. *Social Cognitive & Affective Neuroscience, 2*(1), 39-44.

Rotshtein, P., Henson, R. N., Treves, A., Driver, J., & Dolan, R. J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nature Neuroscience, 8*(1), 107-113.

Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in Black and White children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology, 39*(4), 590-598.

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology, 16*(2), 235-239.

Saxe, R. (2009). Theory of mind (neural basis). In W. P. Banks (Ed.), *Encyclopedia of Consciousness* (Vol. 2, pp. 401-410). Oxford, UK: Elsevier.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage, 19*(4), 1835-1842.

Schiller, D., Freeman, J. B., Mitchell, J. P., Uleman, J. S., & Phelps, E. A. (2009). A neural mechanism of first impressions. *Nature Neuroscience, 12*(4), 508-514.

Schneider, D. J. (2004). *The psychology of stereotyping*. New York City: Guilford Press.

Simmons, W. K., & Martin, A. (2009). The anterior temporal lobes and the functional architecture of semantic memory. *Journal of the International Neuropsychological Society, 15*(5), 645-649.

Simmons, W. K., Reddish, M., Bellgowan, P. S., & Martin, A. (2010). The selectivity and functional connectivity of the anterior temporal lobes. *Cerebral Cortex, 20*(4), 813-825.

Slotnick, S. D., Moo, L. R., Segal, J. B., & Hart, J., Jr. (2003). Distinct prefrontal cortex activity associated with item memory and source memory for visual shapes. *Cognitive Brain Research, 17*(1), 75-82.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Spears, R., Oakes, P. J., Ellemers, N., & Haslam, S. A. (1997). Introduction: The social psychology of stereotyping and group life. In R. Spears, P. J. Oakes, N. Ellemers & S. A. Haslam (Eds.), *The social psychology of stereotyping and group life* (pp. 1-19). Malden, MA: Blackwell Publishing.

Todorov, A. B., Fiske, S. T., & Prentice, D. A. (2011). *Social neuroscience: Toward understanding the underpinnings of the social mind*. New York City, NY: Oxford University Press.

Van Bavel, J. J., & Cunningham, W. A. (2009). A social neuroscience approach to intergroup perception and evaluation. In W. P. Banks (Ed.), *Encyclopedia of Consciousness* (pp. 379-388). New York City: Academic Press.

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*, 237-284.

Wagner, D. D., Kelley, W. M., & Heatherton, T. F. (2011). Individual differences in the spontaneous recruitment of brain regions supporting mental state understanding when viewing natural social scenes. *Cerebral Cortex, 21*(12), 2788-2796.

Wattenmaker, W. D. (1995). Knowledge structures and linear separability: Integrating information in object and social categorization. *Cognitive Psychology, 28*(3), 274-328.

Waytz, A., & Young, L. (2012). The group-member mind trade-off: Attributing mind to groups versus group members. *Psychological Science, 23*(1), 77-85.

Weil, R. S., & Rees, G. (2010). Decoding the neural correlates of consciousness. *Current Opinion in Neurology, 23*(6), 649-655.

Wheatley, T., Milleville, S. C., & Martin, A. (2007). Understanding animate agents: Distinct roles for the social network and mirror system. *Psychological Science, 18*(6), 469-474.

Wheeler, M. E., & Fiske, S. T. (2005). Controlling racial prejudice: Social-cognitive goals affect amygdala and stereotype activation. *Psychological Science, 16*(1), 56-63.

Wilson, E. O. (2012). *The social conquest of earth* (1st ed.). New York: Liveright.

Wong, C., & Gallate, J. (2012). The function of the anterior temporal lobe: A review of the empirical evidence. *Brain Research, 1449*, 94-116.

Xu, X., Yue, X., Lescroart, M. D., Biederman, I., & Kim, J. G. (2009). Adaptation in the fusiform face area (FFA): Image or person? *Vision Research, 49*(23), 2800-2807.

Yzerbyt, V., & Demoulin, S. (2010). Intergroup relations. In D. T. Gilbert, S. T. Fiske & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 2, pp. 1084-1121). New York City: Wiley.

Zahn, R., Moll, J., Krueger, F., Huey, E. D., Garrido, G., & Grafman, J. (2007). Social concepts are represented in the superior anterior temporal cortex. *Proceedings of the National Academy of Sciences United States of America, 104*(15), 6430-6435.

Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and "false" photographs. *Cognition, 35*(1), 41-68.

Zink, C. F., Tong, Y., Chen, Q., Bassett, D. S., Stein, J. L., & Meyer-Lindenberg, A. (2008). Know your place: Neural processing of social hierarchy in humans. *Neuron*, *58*(2), 273-283.