UNIVERSITY OF CALIFORNIA

Santa Barbara

Top-down influences of mere group membership on face representations: The roles of ingroup positivity, category labels, and the self

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Psychological and Brain Sciences

by

Youngki Hong

Committee in charge:

Professor Kyle Ratner, Chair

Professor Daniel Conroy-Beam

Professor Barry Giesbrecht

Professor Diane Mackie

June 2021

The dissertation of Youngki Hong is approved.

Daniel Conroy-Beam

Barry Giesbrecht

Diane Mackie

Kyle Ratner, Committee Chair

June 2021

Top-down influences of mere group membership on face representations: The roles of ingroup positivity, category labels, and the self

ACKNOWLEDGEMENTS

[First, I would like to express my sincere and heartfelt appreciation toward my advisor, Dr. Kyle Ratner, for his support, guidance, and patience over the past six years. I feel deeply indebted and thankful to him for many, if not all, of my accomplishments throughout my graduate career.

I would also like to thank my current and former lab mates – Dr. Locke Welborn, Amanda Kaczmarek, and Diego Padilla-Garcia as well as my colleagues and friends outside of our lab – Evan Layher, Anudhi Munasinghe, and Dipanjana Das for their academic and emotional support. They helped me become not only a better scientist but also a better person. I feel very fortunate to have built life-long friendships with them.

I also want to thank Drs. Diane Mackie, Nancy Collins, Angela Maitner, Daniel Conroy-Beam, and Barry Giesbrecht for providing valuable feedback and perspectives on my work as well as their support during my job search.

Last but not least, I would like to thank my parents for all their support and guidance as well as countless sacrifices they have made to help me get to where I am today, not just during my time in graduate school but my whole life. I would not be where I am today without them.]

Abbreviated CV of Youngki Hong

EDUCATION

University of California, Santa Barbara 2015 – 2021
Doctor of Philosophy in Social Psychology

University of Minnesota, Twin Cities 2011 – 2015
Bachelor of Science in Psychology
Bachelor of Arts in Statistics
Graduation with Summa Cum Laude and High Distinction

PUBLICATIONS

Hong, Y., & Ratner, K. G. (2021). Minimal but not meaningless: Seemingly arbitrary category labels can imply more than group membership. *Journal of Personality and Social Psychology, 120(3)*, 576-600.

Welborn, B. L., **Hong, Y**., & Ratner, K. G. (2020). Exposure to negative stereotypes influences the representations of monetary incentives in the nucleus accumbens. *Social Cognitive and Affective Neuroscience, 15(3)*, 347-358.

Hu, C., Yin, J., Lindenberg, S., Dalgar, I., Weissgerber, S. S., Vergara, R. C., Cairo, A. H, Čolic, M. V., Dursun, P., Frankowska, N., Hadi, R., Hall, C. J., **Hong, Y**., …, & IJzerman, H. (2019). Data from the Human Penguin Project: A cross-national dataset testing principles from social thermoregulation theory. *Scientific Data, 6(1)*, 32.

Ratner, K. G., Kaczmarek, A. R., & **Hong, Y**. (2018). Can over-the-counter pain medications influence our thoughts and emotions? Policy Insights from the *Behavioral and Brain Sciences, 5(1),* 82-89.

IJzerman, H., Dalgar, I., Weissgerber, S. S., Vergara, R. C., Cairo, A. H, Čolic, M. V., Dursun, P., Frankowska, N., Hadi, R., Hall, C. J., **Hong, Y**., …, & Lindenberg, S. M. (2018). The human penguin project: Complex social integration buffers human core temperatures from cold climates. *Collabra: Psychology, 4(1)*, 37.

IJzerman, H., Čolic, M. V., Hennecke, M., **Hong, Y**., …, & Lindenberg, S. M. (2017). Does distance from the equator predict self-control? Lessons from the Human Penguin Project. *Behavioral and Brain Sciences, 40*.

SELECT PRESENTATIONS

Hong, Y., Mayes, M. S., Munasinghe, A. P., & Ratner, K. G. (April, 2021). Neural responses to low-level visual indicators of minimal ingroup and outgroup faces. Presented at the annual meeting of the Social and Affective Neuroscience Society, Virtual Meeting.

Hong, Y., & Ratner, K. G. (February, 2020). The role of ambiguity in intergroup face perception. Presented at Social Cognition preconference of the Society for Personality and Social Psychology, New Orleans, LA.

Hong, Y., Maitner, A., & Ratner, K. G. (February, 2020). Exposure to political rhetoric during the 2016 elections shifted American and Arab people's mental representations of each other. Presented at the annual meeting of the Society for Personality and Social Psychology, New Orleans, LA.

Hong, Y., & Ratner, K. G. (February, 2019). Overestimators are not underestimators: Novel category labels are meaningful when visualizing ingroup and outgroup faces. Presented at the annual meeting of the Society for Personality and Social Psychology, Portland, OR.

Hong, Y. (May, 2018). The relationship between core body temperature and intergroup bias. Paper presented at the annual meeting of Association for Psychological Science, San Francisco, CA.

SELECT TEACHING

University of California Santa Barbara

Instructor of Record

Social Psychology	Summer 2018
Statistics	Winter 2019
Social Cognition	Summer 2019

HONORS/AWARDS

UCSB Graduate Division Dissertation Fellowship
Charles G. McClintock Award
SPSP Graduate Travel Award
Kavli Summer Institute in Cognitive Neuroscience Fellowship
SANS Poster Award
UCSB GSA Excellence in Teaching Award nominee (x2)
Phi Beta Kappa

PROFESSIONAL SERVICES

Ad-hoc Review for Journal of Experimental Social Psychology (x2)
Ad-hoc Review for Nature Human Behaviour
Ad-hoc Review for Social Cognitive and Affective Neuroscience (x3)
Ad-hoc Review for Social Psychological and Personality Science
SPSP Poster Award Review

PROFESSIONAL AFFILIATIONS

Association for Psychological Science (APS)
Social and Affective Neuroscience Society (SANS)
Society for Personality and Social Psychology (SPSP)
Society for Psychophysiological Research (SPR)

ABSTRACT

Top-down influences of mere group membership on face representations: The roles of ingroup positivity, category labels, and the self

by

Youngki Hong

[How do people determine what an ingroup looks like? Past research using a minimal group paradigm suggests that people imbue ingroups with physical features that convey desirable attributes. In this research, I used the classic overestimator versus underestimator and Klee versus Kandinsky minimal group paradigms and the reverse correlation method to examine various top-down influences of mere group membership on face representations of ingroup and outgroup, beyond the well-documented ingroup positivity effect. In Study 1a, I show that participants represented ingroup faces more favorably than outgroup faces, but also represented faces of overestimator and underestimator groups differently. In fact, the category label effect was larger than the ingroup positivity effect. In Study 1b, I demonstrate that faces of Klee and Kandinsky groups were also represented differently at the group-level, but not at the participant-level. This lack of category label effect was in turn related to a stronger ingroup positivity effect. In Study 2a and 2b, I show that people were more likely to use their own self-image to mentally represent ingroup faces than outgroup faces. In Study 3a and 3b, I show that self-evaluation was related to the extent to which ingroup positivity bias

was expressed in people's mental representation of ingroup and outgroup faces, more so than group-evaluation, and also provide a potential alternative explanation for the findings from Study 2. Together, this work advances but does not upend understanding of minimal group effects. I robustly replicate the ingroup positivity effect in face representations. At the same time, I demonstrated other top-down influences of such as category labels as well as the self-knowledge play on how people visually represent faces of minimal ingroup and outgroup members.]

Keywords: Minimal Group Paradigm, Faces, Reverse Correlation, Machine Learning, Representational Similarity Analysis, Self

Table of Contents

Top-down influences of group membership on face representations: The roles of ingroup positivity, category labels, and the self

Introduction

Fifty years ago, Tajfel, Billig, Bundy, and Flament published a now classic article that launched the Minimal Group Paradigm (1971). The minimal group paradigm challenged the conventional wisdom that discrimination requires conflict over resources or animosity between groups. Instead, merely separating people into arbitrary groups creates a variety of intergroup biases. People often have extended knowledge about real-world groups, such as stereotypes and life experiences with members of the groups, all of which can guide their judgments and behaviors in intergroup contexts (Kunda & Spencer, 2003). The novelty of the categories in the minimal group paradigm is a celebrated feature of the paradigm because it is thought to strip away the complexity that makes established group distinctions, such as race and gender, so difficult to study. For instance, when investigating the dynamics that contribute to race bias, it is often challenging to know whether effects are due to status or power differences between the groups (Weisbuch et al., 2017), stereotypes circulating in the culture (Devine, 1989), personal antipathy (Augoustinos & Rosewarne, 2001), direct experience with the groups (Columb & Plant, 2016; Qian et al., 2017), own group preference (Lindström et al., 2014), or other confounding variables. Tajfel and colleagues (1971) cleverly devised experimental paradigms to sidestep this problem by using novel group categories to circumvent the influences of preexisting knowledge people have about real-world groups. Novel group categories by virtue of their novelty signal whether a target shares one's group membership or not, but do not convey much else beyond that.

Intergroup bias in face processing

Mere group categorization into minimally defined groups can influence a wide range of responses, including perception, attitudes, emotions, and behaviors in favor of one's ingroup, a phenomenon known as ingroup favoritism (Ashburn-Nardo, Voils, & Monteith, 2001; Brewer & Silver, 1978; Brown, & Abrams, 1986; Dunham, Baron, & Carey, 2011; Howard & Rothbart, 1980; Locksley, Ortiz, & Hepburn, 1980; Otten & Moskowitz, 2000; Tajfel et al., 1971). The way we perceive and process faces is not free from the effects of minimal group memberships; abundant research on face processing in minimal group context exists in the literature (Bernstein, Young, & Hugenberg, 2007; Hugenberg & Corneille, 2009; Lazerus et al., 2016; Ratner & Amodio, 2013; Ratner et al., 2014; Van Bavel, Packer, & Cunningham, 2008, 2011; Young & Hugenberg, 2010). Although intergroup biases in minimal group context have been studied in many domains, there are a number of reasons for focusing on face processing in particular. First, many important interactions occur face-to-face, both interpersonal and intergroup (Maclnnis & Page-Gould, 2015; Macrae et al., 2005; Oosterhof & Todorov, 2008; Todorov et al., 2008); one of the most classic examples being prejudice reduction through direct contact with outgroup members (Allport, 1954; Pettigrew, 1998; Pettigrew & Tropp, 2008). Second, due to the amount and richness of information conveyed by faces, different cognitive and motivational processes are involved in face processing (Axelrod, Bar, & Rees, 2015). Therefore, the use of faces provides a unique opportunity for exploring multiple aspects of intergroup behaviors. For example, people easily make inferences about other people's characters from faces (Cogsdill et al., 2014; Hassin & Trope, 2000; Oosterhof & Todorov, 2008; Todorov et al., 2008), which often lead to important social outcomes, such as election (Todorov et al., 2005) and criminal sentencing

(Blair, Judd, & Chapleau, 2004; Eberhardt et al., 2006). Thus, focusing in on how people process faces of ingroup and outgroup members can be crucial in understanding intergroup biases and conflicts.

Despite the widely held interest, the nature of intergroup face processing is not well understood. One major difficulty in studying intergroup face processing is that the downstream effects of group membership on face processing we often see in the literature could be based on a number of both top-down and bottom-up factors[1]. That is, it can be ambiguous whether the effects are due to top-down influences, bottom-up information, or some combination of both. On the one side, as is the case for any visual perception, face processing is highly sensitive to low-level visual features such as luminance, contrast, and spatial frequency of visual stimuli (i.e., faces) (Weibert et al., 2018). Going beyond low-level visual features, specific facial features are reliably associated with specific trait judgments (Oosterhof & Todorov, 2008). On the other side, there exist numerous top-down influences on face processing, including ingroup positivity (Ratner et al., 2014) and stereotypes and attitudes (Freeman & Johnson, 2016), and task goals (Kaul, Ratner, & Van Bavel, 2014). Because of such complexities involved in intergroup face processing, social psychologists have started using a method called the reverse correlation methods to study visual representation[2] people might have about various groups (see Brinkman, Todorov, & Dotsch, 2017; Dotsch & Todorov, 2012 for review). By studying visual imagery, researchers remove

[1] I define top-down factors as any preexisting information or motivation people have that influence subsequent processing of visual stimuli (Gilbert & Yi, 2013), whether consciously exerted by the perceiver or not.

[2] By representation, I mean a visual approximation of the content of mental imagery people might have about various groups. This interpretation is similar to how fMRI researchers view the BOLD signal as an approximation of neural responses.

the influence of any meaningful bottom-up information and isolate the top-down influence, and thus have a better handle for the effect of top-down influences on face processing.

Of particular relevance to the current research, Ratner and colleagues (2014) used the reverse correlation method to demonstrate that people visualize minimal ingroup faces differently than minimal outgroup faces. By using the reverse correlation method and the minimal group paradigm, the researchers removed not only the influence of meaningful bottom-up information but also preexisting knowledge people might have about real-world groups. This mere group categorization led to only the motivational influence of group membership – the motivation to view ingroup more favorably than outgroup (Tajfel & Turner, 1979) – as indicated by more favorable face representation of ingroup than outgroup. However, does mere categorization really remove all top-down influences of knowledge that people can bring to bear to make sense of groups? In my dissertation research, I explored top-down influences of knowledge structures on people's face representations of ingroup and outgroup members even we try to remove their influences using the minimal group paradigm.

Top-down influences of knowledge structures on face representations

Category labels. It is no coincidence that Tajfel et al. (1971) used rather contrived group distinctions, overestimators versus underestimators and preference for paintings by Klee versus Kandinsky, in their landmark article. Much of the minimal group research that followed used paradigms that implied similarity of novel group members (as was the case with Tajfel's contrived group distinctions and also the use of personality tests, e.g., Bernstein et al., 2007) or common fate implied by a competitive context (e.g., Cikara et al., 2014; Van Bavel & Cunningham, 2009) to create entitative groups. The qualifier minimal in the name of the paradigm reflects the fact that there is typically not a complete absence of differences

between groups. Two years prior to the inception of the minimal group paradigm, Rabbie and Horwitz (1969) reported that when group distinctions are completely arbitrary then impressions of novel ingroup and outgroup members did not differ. Although Billig and Tajfel (1973) later reported that overt random assignment could lead to biases in resource allocations, the effect of this random grouping on intergroup bias was much weaker than was the case with their original contrived group distinctions. Thus, in a typical minimal group situation, participants are confronted with categories that are plausible but novel to them. From the perspective of a participant in a minimal group situation, they are faced with a task (e.g., making judgments of faces, allocating resources, evaluating people), but the experimenter has made this task difficult by putting them in an explanatory vacuum. Without this concrete knowledge, it is assumed by researchers that participants will default to a heuristic that ingroups should be preferred, for example, to maintain self-esteem (Hogg & Abrams, 1990; Tajfel & Turner, 1979). These perspectives see the category label merely as a mechanism to signal who is an ingroup member and who is an outgroup member and that labels themselves do not carry much meaning. However, it is notable that in most cases the category label distinction (e.g., overestimator versus underestimator) is explicit and the ingroup versus outgroup distinction is implicit (e.g., allocating points to an overestimator and an underestimator person *not* an ingroup and an outgroup member in Tajfel et al., 1971). This overt emphasis on the category labels gives participants several reasons to not behave according to the experimenter's desires and instead try to make sense of the minimal group labels and use them to guide their task decisions.

Social perception research finds that when perceivers find themselves contemplating puzzling situations, such as when a target person is described as having conflicting trait

attributes or a target person performs behaviors that are incongruent with an existing schema, perceivers engage in reasoning to make sense of what confuses them (Asch & Zukier, 1984; Hastie, 1984). This type of processing is consistent with a long tradition in psychology of characterizing people as meaning-makers, including Bruner's observation that people frequently go beyond the information given (Bruner, 1957), work on epistemic motives suggesting that people have a need to understand the world around them (Cacioppo & Petty, 1982; Kruglanski & Webster, 1996), neuropsychology research on the tendency for people to confabulate to make sense of confusing circumstances (Gazzaniga, 2000), early social cognition work suggesting that people go about telling more than they can know (Nisbett & Wilson, 1977), and social–cognitive models of transference that show that when novel targets superficially resemble a significant other, trait information about the significant other is applied to make sense of the novel person (Andersen & Cole, 1990).

In the puzzling explanatory vacuum that is the minimal group paradigm, the category labels might provide knowledge structure latch onto that can fill in inferential gaps. Although people might not have experiences with the category labels, they might come up with different associations that give them meaning. For instance, consider the popular overestimator/underestimator minimal group paradigm from the original Tajfel and colleagues' studies (1971). The label overestimator is objectively defined (in an experimental situation) as someone who assumes there is more of a quantity than is actually the case. Who is likely to be an overestimator? Maybe people who are optimistic or confident or people who are arrogant. Underestimators, on the other hand, might be more cautious and timid. These inferences can lead to assumptions about who is more dominant and who you should

be more likely to trust to not take advantage of you. In this way, perceivers can very quickly go beyond the information given.

Self-as-a-representational base hypothesis. Prominent social groupings, such as family units, friend networks, racial categories, and ethnic tribes are often bonded through a shared cultural and genetic heritage that results in members of a group sharing appearance cues ranging from how they dress and cut their hair to the contours and features of their faces (DeBruine, 2004; Hehman, Flake, & Freeman, 2018). This suggests that people may infer that people who look like them are part of their group. However, is the notion that ingroup members physically resemble the self so ingrained in social category representation that it persists even when groups have no shared social history and are not defined by physical attributes? The only non-ambiguous information participants know about the minimal group paradigm is that they belong to their ingroup. Therefore, if participants have to visualize what their ingroup looks like they could use their self-image as a basis for their visualization. Ingroup positivity bias could then be observed in ingroup and outgroup classification images because people tend to have self-serving biases (Feingold, 1992) so the self-image they use as a basis for deciding what an ingroup member looks like could be distorted by rose-colored glasses. I call this possibility the *self-as-a-representational-base hypothesis*. This self-as-a-representational bias hypothesis would be consistent with work by Ames (2004a, 2004b) showing that perceived similarity to a target triggers self-projection. It also fits with the self-as-an-informational base (Gramzow, Gaertner, & Sedikides, 2001) and self-as-an-evaluative base (Gramzow & Gaertner, 2005) frameworks that have been proposed to understand ingroup bias in memory and evaluation in minimal group situations.

Self-as-an-evaluative-base hypothesis. Closely related to the self-as-a-representational base hypothesis, people may simply use how they feel about themselves to guide their decisions in the face categorization task. People not only base their self-worth on their group memberships and social identities (Tajfel & Turner, 1979), but also base their judgments of their groups on how they evaluate themselves (Hogg & Abrams, 1990). Thus, when people regard themselves highly, they may also evaluate their ingroup more positively. This was in fact demonstrated by Gramzow and Gaertner (2005) using the minimal group paradigm in which they found that personal self-esteem was related to how people rated their novel ingroup and outgroup (*the self-as-an-evaluative-base hypothesis*). It is possible that how people evaluate themselves would also be related to how favorable or unfavorable their face representations of ingroup and outgroup look. Ingroup positivity bias would be observed because most people evaluate themselves positively (Taylor & Brown, 1988).

Overview of the current research

The current research was designed to examine three primary objectives. The first was to determine if people imbue classic minimal group labels with any meaning whatsoever above and. The second was to examine whether the inductive potential of the labels is critical levers on whether category labels have an influence on mental representation of minimal ingroup and outgroup faces. The last objective was to examine the role of the self in shaping people's mental representation above and beyond the common ingroup positivity effect. To address these aims, I turned to reverse correlation image classification, which is a technique that has been used widely in social psychology to investigate how people represent social categories (for a review, see Brinkman et al., 2017). As mentioned above, Ratner et al. (2014) investigated how minimal ingroup and outgroup faces are represented. They used

reverse correlation to show that ingroup and outgroup faces are represented differently, but their analyses did not address whether category labels were also represented differently. In Study 1a, I conducted a preregistered, highly powered replication of Ratner et al.'s (2014) Study 1 to first replicate their ingroup positivity findings. I go beyond Ratner et al. (2014), however, by also examining representational differences between overestimator and underestimator. This latter analysis provides insight into whether the overestimator versus underestimator distinction was represented, which would be consistent with participants inferring meaning from the category labels. I also use representational similarity analysis (Kriegeskorte, Mur, & Bandettini, 2008; Stolier, Hehman, & Freeman, 2018) to examine whether the overestimator/underestimator distinction or the ingroup/outgroup distinction equally contribute to the face representations or if one distinction is weighted to a greater degree than the other. Lastly, I used a machine learning approach to examine the representational differences between ingroup and outgroup as well as overestimator and underestimator at the participant-level to provide convergent support. Study 1b examines the generalizability of these phenomena by replicating the findings from Study 1a with the Klee versus Kandinsky minimal group paradigm, a version of the minimal group paradigm that uses labels that less clearly imply traits of the group members than does the overestimator versus underestimator paradigm (Tajfel et al., 1971). In Study 2, I tested the self-as-a-representational base hypothesis. I did so by using the participant-level classification images and photographs of people who generated those images from Study 1 to test the idea that people might have used their own self-image to visually represent faces of their ingroup but not outgroup. Study 3 went beyond the use of self-image in people's face representation of ingroup and outgroup and tested whether people's self-evaluation (self-esteem) is related to

the perceived trustworthiness of ingroup and outgroup images they generated in an economic trust game. The current research makes many important theoretical contributions to our understanding of not only the minimal group paradigm but also various top-down influences of group membership on face representations. First, by using minimal group paradigms that have varying degrees of inductive potential, the current research tests top-down influences of these labels on people's face representations of ingroup and outgroup and challenges the assumption that Tajfel and many researchers who followed him made about minimal group labels—that they are flimsy and uninformative. Second, by examining the role of the self in the face representations, I highlight the importance of examining multiple sources to top-down influences on face representations. In order to address these questions, I also introduce methodological innovations that could be helpful to other researchers who use reverse correlation methods. I draw on methods and analytic techniques commonly used in cognitive and computational neuroscience, such as machine learning and representational similarity analysis to shed light on processes and top-down influences that could be difficult to study using conventional analytic tools.

Study 1 – the category label effects

Study 1a

There were multiple goals of Study 1a. First was to replicate the findings of Ratner et al. (2014) by demonstrating that people show ingroup positivity in face representations. Second, I tested for differences in people's face representations of overestimators and underestimators. Third, after establishing that minimal group labels might have meaningful distinctions in face representations, I sought to understand how the ingroup/outgroup distinction and the overestimator/underestimator distinction differentially contribute to face

representations. Lastly, I examined whether the representational differences between overestimator and underestimator exist at the participant level using a novel method of analyzing reverse correlation images with machine learning. To clearly demarcate between different approaches to analyzing data, I break down Study 1a into three parts (analysis 1,2, and 3).

Analysis 1

Data collection for Study 1a and specifically Analysis 1 was conducted in two phases, in which I (a) created visual renderings of participants' face representations of minimally defined groups, and (b) collected trait ratings of these images from a separate group of participants naïve to the face generation stage. In Phase 1, participants were randomly assigned to minimal groups and then categorized faces as belonging to either of two minimal groups. I used the reverse correlation image classification technique to create mental representations of faces. The reverse correlation method examines response biases to different stimuli to infer patterns in the stimuli that may have caused the responses. These patterns then are visualized and provide an approximation of the mental representation upon which participants based their responses (Dotsch & Todorov, 2012). In Phase 2, I assessed whether these visual renderings could reveal differences in face representations of different groups by asking an independent sample of participants to rate classification images of different minimal group faces. Following Ratner et al. (2014), the faces were rated on 13 trait dimensions that Oosterhof and Todorov (2008) used to assess trait impressions of faces.

Analysis 1 was designed to determine whether people have different mental representations of faces of different minimal groups. On the one hand, overestimator and underestimator distinctions might not be represented by the perceiver—these minimal groups

are designed to be arbitrary and novel to the participants and thus, there is no intended basis for a difference other than whether a target shared the same group membership (ingroup) or not (outgroup). On the other hand, people are motivated to make sense of the world around them (Cacioppo & Petty, 1982; Kruglanski & Webster, 1996) as evident from cases of confabulation in patients (Gazzaniga, 2000) and nonpatients (Nisbett & Wilson, 1977), and social–cognitive transference (Andersen & Cole, 1990). Thus, they could imbue novel category labels with meaning, infer different traits from them, and generate different face representations as a result. Thus, I remained agnostic a priori about whether the minimal group labels would have an influence on face representations. I predicted that people would show ingroup positivity as indicated by more favorable trait ratings of ingroup faces than outgroup faces.

Method

Participants. I recruited 362 University of California, Santa Barbara students (M_{age} = 18.92, *SD* = 1.61; 245 female, 109 male, and eight unidentified) to participate in a study about categorizing faces in exchange for course credit via the UCSB Psychological and Brain Sciences subject pool. I sought to maximize our power by more than doubling the sample size of a similar study that utilized the same procedure and methods (*n* = 174 of Study 1 from Ratner et al., 2014). The racial and ethnic breakdown of our sample was 110 White, 106 Latinx, 90 Asian, 22 multiracial, two Pacific Islander/ Hawaiian, 15 other, and eight unidentified. Up to six participants were run simultaneously.

Procedure. As with Ratner et al. (2014), participants were first told that they would perform several tasks on a computer. Next, the Numerical Estimation Style Test (NEST) version of the classic "dot estimation" procedure (Experiment 1 from Tajfel et al., 1971) was

used to assign participants to novel, but believable, groups (Ratner & Amodio, 2013; Ratner et al., 2014). Then participants completed a face categorization task optimized for a reverse correlation analysis.

Numerical Estimation Style Test (NEST). In this task, participants were told that people vary in numerical estimation style, which was defined as the tendency to overestimate or underestimate the number of objects they encounter. I also told the participants that approximately half the population are overestimators and half are underestimators, and that there is no relationship between numerical estimation style and any other cognitive tendencies or personality traits.1 I then told participants that they would categorize photographs of students from a previous quarter whose numerical estimation style had been determined with a well-established task called the Numerical Estimation Style Test (NEST). I also told them that people can reliably detect numerical estimation style from faces and that the purpose of the study was to test whether people can determine numerical estimation style when faces appear blurry.

Next, participants completed the NEST themselves. In this task, they attempted to estimate the number of dots in 10 rapidly presented dot patterns, which each appeared for 3,000 ms. At the end of the test, the computer program provided predetermined feedback (counterbalanced across participants), indicating that each participant was either an overestimator or underestimator. I did not actually take participants' NEST responses into account; the NEST was used to provide a rationale for the group assignment.

I used additional procedures to make the novel groups (i.e., overestimator and underestimator) as salient as possible in participants' minds throughout the remainder of the study. First, participants reported their numerical estimation style to the experimenter,

providing a public commitment to their ingroup. The experimenter then wrote each participant's identification number and numerical estimation style on a sticky note and attached it to the bottom center of the computer monitor (in the participants' line of sight) to constantly remind them of their group membership during the face categorization task. Participants also typed their numerical estimation style into the computer, as another act of commitment to the ingroup.

Face categorization. After the group assignment, participants completed a forced-choice face categorization task for 450 trials. On each trial, participants selected either an overestimator or underestimator face out of two adjacent grayscale face images. Half of the participants were asked on every trial to choose which of the two faces was an overestimator and the other half of the participants were asked on every trial to choose underestimator faces. If the targets shared the same numerical estimation style (i.e., overestimator or underestimator) with the participant, then the participant was selecting ingroup faces, whereas if the targets did not share the same numerical estimation style with the participant, then the participant was selecting outgroup faces.

I used the grayscale neutral male average face of the Averaged Karolinska Directed Emotional Faces Database (Lundqvist, Flykt, & Öhman, 1998) as the base image to generate 450 pairs of face stimuli used in the face categorization task. Different noise patterns, which consisted of 4,092 superimposed truncated sinusoid patches, were added to the same base image, generating 450 different face pairs (Dotsch & Todorov, 2012; Mangini & Biederman, 2004; Ratner et al., 2014). A noise pattern was applied to the base image, and the inverse of that noise pattern was added to the base image, creating a pair of images. I presented inverse

noise faces equally on the left and right sides of the screen in a random order. I used the same pairs of faces for all participants.

Individual differences questionnaire. After the face categorization, participants completed an individual differences questionnaire packet that included scales for personal self-esteem (PSE) and collective self-esteem (CSE) administered through Qualtrics (www.qualtrics.com). Both scales yielded reasonable Cronbach's α levels: PSE = .87 and CSE = .85. The scale data are analyzed in Study 3.

Photographs. Finally, I gave them a separate consent form asking for their permission to have their photographs taken. I also informed them that their photographs would be used in our future studies. I asked our participants to maintain a neutral facial expression to match the base image used in the study. I did not penalize anyone who declined to consent to have their photograph taken. The photographs were used in a separate study and the findings are presented in Study 2.

Face representation data processing. Following the logic of reverse correlation analysis, I generated visual renderings of different groups by averaging noise patterns of selected faces (Dotsch & Todorov, 2012; Dotsch et al., 2008; Dotsch, Wigboldus, & van Knippenberg, 2011). I argue that the reverse correlation analysis is suitable for capturing the difference between overestimator and underestimator face representations because if participants selected faces based solely on their group membership, overestimator and underestimator faces should look the same. If participants imbued meaning into the category labels, then systematic patterns would reveal the difference in mental representations of overestimator and underestimator faces. Thus, using the reverse correlation method allowed

for examining not only biases in favor of the ingroup, but also differences between overestimator and underestimator face representations.

Participant-level classification images. The R package, *rcicr* (Dotsch, 2016), was used to conduct the reverse correlation analysis. I first averaged noise patterns of the chosen 450 faces from the face categorization task for each participant and superimposed the normalized average noise pattern back onto the original base image to create participant-level classification images. The images reflected participants' mental representations of what an overestimator or underestimator face should look like. A classification image was ingroup if the target's group membership was shared with that of the participant, whereas the image was outgroup if the target's group membership was different from that of the participant.

Group-level classification images. After creating participant-level classification images, I created eight group-level classification images. First, to test whether I replicated the ingroup positivity effect found in Study 1 of Ratner et al. (2014), I created ingroup (n = 180) and outgroup (n = 182) classification images by averaging the appropriate noise patterns from the participant-level. That is, I averaged noise patterns of participant-level classification images of ingroup faces and super- imposed the normalized average noise pattern back onto the base image to create the group-level classification image for the ingroup face. I did the same for the outgroup face (see Figure 1). Second, to examine the difference in trait impressions elicited by the category labels, I also created overestimator (n = 181) and underestimator (n = 181) classification images by following the same procedure for the ingroup and outgroup group-level classification images (see Figure 2). Finally, I examined the interaction between group membership and the category labels by creating four classification images by crossing the two variables: ingroup-overestimator (n = 91), ingroup-

underestimator (n = 89), outgroup-overestimator (n = 90), and outgroup-underestimator (n = 92). All four classification images can be seen in Figure 3.

Figure 1. Study 1a ingroup and outgroup group-level classification images

Figure 2. Study 1a overestimator and underestimator group-level classification images.

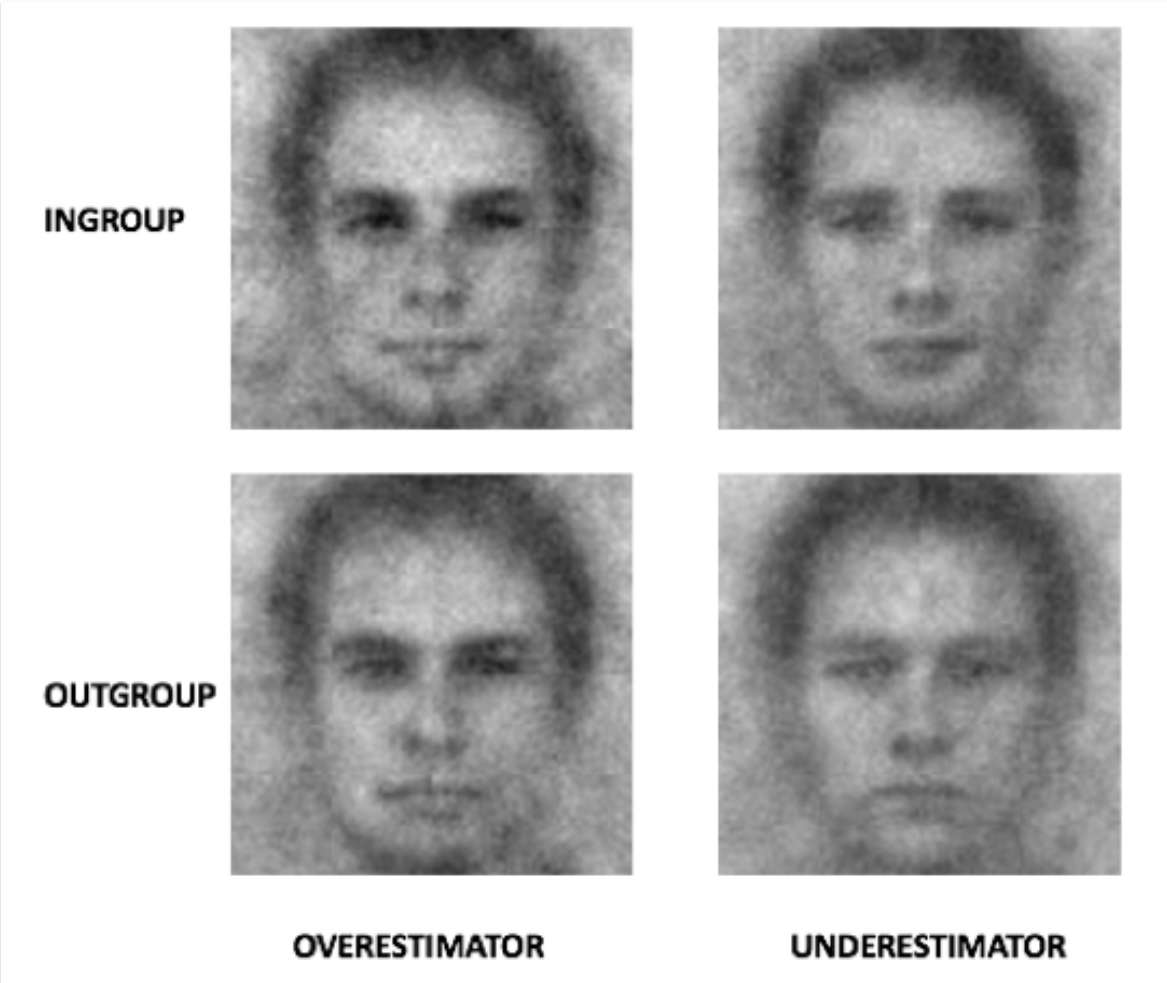


Figure 3. Study 1a GROUP X NEST group-level classification images

Phase 2: Assessing impressions of face representations.

In Phase 2, I objectively assessed the differences in these face representations, specifically in how they elicited different trait impressions. To do this, I had independent samples of participants who were not aware of the face categorization stage from Phase 1 rate the eight group-level classification images from Phase 1. To assess *relative* differences between ingroup and outgroup (Group), overestimator and underestimator (NEST), and Group X NEST images, I obtained ratings from three different samples of participants. That

is, participants only rated ingroup and outgroup images, overestimator and underestimator images, or Group X NEST images.

Participants. I recruited a total of 301 participants (M_{age} = 35.98, *SD* = 11.44; 145 female, 156 male) through the TurkPrime website (www.turkprime.com) to complete an online survey administered through Qualtrics (www.qualtrics.com). Ninety-nine participants rated ingroup and outgroup classification images, 102 participants rated overestimator and underestimator classification images, and 100 participants rated ingroup-overestimator, outgroup-overestimator, ingroup-underestimator, and outgroup-underestimator classification images. I recruited a comparable number of Mechanical Turk (MTurk) raters as Ratner et al. (2014). The racial and ethnic breakdown of the raters was 226 White, 28 Asian, 22 Black, and 11 multiracial participants. The samples in this portion of the study were collected from MTurk, which are comparable with typical undergraduate student samples, if not more diverse (Buhrmester, Kwang, & Gosling, 2011). Participants were expected to complete the study in 10 min. All participants did not know about the face categorization stage of the study (i.e., Phase 1). They were compensated with $1 for their participation.

Procedure. Participants rated the classification images on 13 trait dimensions (i.e., To what extent is this face . . . trustworthy, attractive, dominant, caring, sociable, confident, emotionally stable, responsible, intelligent, aggressive, mean, weird, and unhappy?; Oosterhof & Todorov, 2008). Each face was presented by itself in a random order. Ratings were made on scales from 1 (*not at all*) to 7 (*extremely*). The order of each trait presentation was also random.

Results

For each sample of raters, I conducted a repeated-measures multivariate analysis of variance (rMANOVA) followed by a univariate analysis of variance for each trait. I show the results below separated by sample.

Group membership (Group). A rMANOVA comparing the trait ratings of ingroup and outgroup classification images was significant, Pillai's Trace = .85, $F = 36.26$, $df = (13, 86)$, $p = .001$, indicating some difference in trait ratings between ingroup and outgroup classification images. The univariate F tests showed that all trait ratings of ingroup and outgroup images were significantly different from each other at the .001 significance level. The means, F values, p values, and effect sizes for each comparison are presented in Table 1. The ingroup face was rated significantly more trustworthy, attractive, caring, emotionally stable, responsible, intelligent, and sociable; the outgroup face was rated significantly more dominant, aggressive, mean, weird, and unhappy.

	Ingroup mean (SD)	Outgroup mean (SD)	F-value	Cohen's d
Trustworthy	4.81 (1.11)	2.94 (1.09)	155.92***	1.25
Attractive	4.51 (1.31)	3.03 (1.18)	109.51***	1.05
Dominant	3.42 (1.36)	4.3 (1.63)	17.73***	.42
Caring	5.00 (1.18)	2.77 (1.32)	183.37***	1.36
Confident	4.90 (1.09)	3.55 (1.38)	62.89***	.80
Emotionally stable	5.08 (1.09)	3.15 (1.31)	139.7***	1.19
Responsible	4.81 (1.04)	3.64 (1.15)	64.01***	.80
Intelligent	4.84 (0.91)	3.7 (1.17)	82.07***	.91
Aggressive	2.60 (1.48)	4.7 (1.40)	137.71***	1.18
Mean	2.42 (1.41)	4.8 (1.44)	127.97***	1.14
Weird	2.58 (1.44)	3.84 (1.71)	59.22***	.77
Unhappy	2.46 (1.33)	5.62 (1.26)	304.24***	1.75
Sociable	5.17 (1.16)	2.61 (1.25)	183.35***	1.36

Significance codes: *** <.001 ** <.01 * <.05 $^{+}$<.10

Table 1. Study 1a trait rating ANOVA results – GROUP (NEST) – face representations

Numerical estimation style (NEST). A rMANOVA comparing the trait ratings of overestimator and underestimator classification images was significant, Pillai's Trace = .68, $F = 14.57$, $df = (13, 89)$, $p = .001$, indicating some difference in trait ratings between overestimator and underestimator classification images. The univariate F tests showed that the majority of trait ratings of overestimator and underestimator images were significantly different from each other at the .001 significance level. The means, F values, p values, and effect sizes for each comparison are presented in Table 2. The overestimator face was rated significantly more dominant, confident, emotionally stable, aggressive, mean, and sociable; the underestimator face was rated significantly more trustworthy, caring, and unhappy.

Attractive, responsible, intelligent, and weird ratings were not significantly different between overestimator and underestimator images.

	Overestimator mean (SD)	Underestimator mean (SD)	F-value	Cohen's d
Trustworthy	3.90 (1.29)	4.40 (1.12)	10.24***	.32
Attractive	4.34 (1.24)	4.28 (1.10)	.15	.04
Dominant	5.10 (1.19)	2.79 (1.32)	124.97***	1.11
Caring	3.70 (1.36)	4.41 (1.36)	15.40***	.39
Confident	5.48 (1.19)	2.93 (1.44)	151.70***	1.22
Emotionally stable	4.46 (1.31)	3.75 (1.33)	14.21***	.37
Responsible	4.35 (1.36)	4.39 (1.12)	.07	.03
Intelligent	4.47 (1.17)	4.45 (1.01)	.02	.02
Aggressive	4.44 (1.60)	2.57 (1.54)	71.20***	.84
Mean	3.99 (1.55)	2.68 (1.50)	39.34***	.62
Weird	3.12 (1.73)	2.82 (1.56)	3.52^{+}	.19
Unhappy	3.51 (1.65)	5.46 (1.20)	93.64***	.96
Sociable	4.31 (1.48)	3.51 (1.36)	15.27***	.39

Significance codes: *** <.001 ** <.01 * <.05 $^{+}$<.10

Table 2. Study 1a trait rating ANOVA results – NEST – face representations

Group X NEST. I used rMANOVA to test the effects of Group, NEST, and the interaction between the two on trait ratings. Significant multivariate effects were found for all variables: the effects of GROUP, NEST, and the interaction between the two on trait ratings. Significant multivariate effects were found for all variables: GROUP (Pillai's Trace = .58, F = 30.57, df = (13, 285), p < .0001), NEST (Pillai's Trace = .48, F = 20.49, df = (13, 285), p < .0001), and GROUP X NEST (Pillai's Trace = .10, F = 2.53, df = (13, 285), p = .003). Similar to the GROUP results reported earlier, ingroup faces were rated more trustworthy, attractive, caring, confident, emotionally stable, responsible, intelligent, and sociable,

whereas outgroup faces were rated more dominant, aggressive, mean, weird, and unhappy for both overestimators and underestimators. Interaction effects were found for some traits including attractive, caring, emotionally stable, aggressive, mean, unhappy, and sociable. The univariate F test results including the means, standard deviations, F values, p values, and effect sizes (comparing ingroup and outgroup within overestimator and underestimator) for each trait are presented in Table 3.

	Overestimator			**Underestimator**			**F-values**		
	Ingroup (SD)	Outgroup (SD)	Cohen's d	Ingroup (SD)	Outgroup (SD)	Cohen's d	Group	NEST	Group X Nest
Trustworthy	4.25 (1.20)	3.39 (1.45)	.50	4.43 (1.42)	3.12 (1.30)	.74	74.11***	.13	3.19^{+}
Attractive	4.36 (1.43)	3.88 (1.57)	.34	4.11 (1.33)	3.03 (1.35)	.77	54.77***	27.23***	8.10**
Dominant	4.32 (1.48)	5.03 (1.21)	.39	2.38 (1.24)	3.54 (1.75)	.60	49.49***	166.50***	2.87^{+}
Caring	4.14 (1.35)	3.12 (1.27)	.57	4.76 (1.20)	2.61 (1.34)	1.35	171.60***	.21	21.81***
Confident	5.37 (.94)	4.97 (1.45)	.26	3.27 (1.58)	2.86 (1.42)	.21	10.08**	272.33***	.00
Emotionally stable	4.64 (1.18)	4.00 (1.41)	.44	4.02 (1.48)	2.80 (1.20)	.70	60.97***	58.38***	5.93*
Responsible	4.48 (1.10)	3.95 (1.27)	.33	4.43 (1.34)	3.50 (1.25)	.53	37.16***	4.36*	2.79^{+}
Intelligent	4.70 (1.10)	4.33 (1.16)	.27	4.67 (1.14)	3.90 (1.11)	.55	31.31***	5.10*	3.85^{+}
Aggressive	3.60 (1.56)	4.57 (1.59)	.50	2.07 (1.21)	4.04 (1.76)	.98	108.17***	53.11***	12.51***
Mean	3.10 (1.59)	4.35 (1.67)	.64	2.22 (1.30)	4.35 (1.77)	1.06	141.45***	9.59**	9.59**
Weird	2.48 (1.30)	3.03 (1.67)	.40	2.79 (1.59)	3.44 (1.78)	.37	28.15***	10.13**	.20^{+}
Unhappy	2.97 (1.39)	3.97 (1.54)	.65	4.18 (1.70)	6.03 (.97)	1.02	121.52***	159.97***	10.81**
Sociable	4.70 (1.44)	3.65 (1.40)	.61	4.17 (1.54)	2.35 (1.08)	1.03	129.91***	52.82***	9.35**

Significance codes: *** <.001 ** <.01 * <.05 $^{+}$<.10

Table 3. Study 1a trait rating ANOVA results – GROUP X NEST – face representations

Discussion

In Analysis 1, I investigated the face representations of minimally defined groups. First, I replicated the findings of Ratner et al. (2014): overall ingroup faces elicited more positive trait impressions compared to outgroup faces. The current study was pre-registered and highly powered - twice the sample size of the original study by Ratner et al. (2014), providing strong evidence that ingroup positivity bias in face representation is a replicable effect.

More interestingly, however, I also found that people generated face representations of overestimators and underestimators differently. The overestimator and underestimator faces differed on various traits dimensions that do not necessarily signal favoritism toward one group over the other. Most notably, the underestimator face image was rated as both more trustworthy and more unhappy than the overestimator face image. This finding is contrary to the general assumption in the literature that the difference between minimal group labels is arbitrary (Tajfel et al., 1971). Instead, the current study showed that people might infer different traits from category labels and utilize them when visualizing faces of ingroup and outgroup members. I also found several interaction effects for GROUP X NEST trait rating data indicating that the magnitude of differences between ingroup and outgroup faces were different for overestimator and underestimator (i.e., larger ingroup positivity for underestimator than overestimator for many traits), providing evidence that face representations of overestimator and underestimator are different and can moderate intergroup bias. For instance, overestimators were generally rated as more attractive, emotionally stable, sociable, aggressive, and mean than underestimators, and perhaps this constrained variability between ingroup and outgroup on these trait dimensions for

overestimators, which resulted in stronger ingroup and outgroup differences on these variables for underestimators. Interestingly, for the caring dimensions, there was no NEST main effect, but there was an interaction effect indicating that the ingroup versus outgroup caring effect was larger for underestimators. Additionally, underestimators were generally rated as more unhappy than overestimators, but the group difference was still larger for underestimator on this variable. Together, there seems to be evidence on multiple trait dimensions that the degree to which ingroup and outgroup differences emerge is influenced by the meaning derived from the overestimator and underestimator distinctions.

Analysis 2

In Analysis 2, I used multiple regression representational similarity analysis (RSA) (Kriegeskorte, 2008; Stolier & Freeman, 2016; Stolier, Hehman, & Freeman, 2018) to more directly examine how much participants were weighting the group (ingroup or outgroup) versus NEST category labels (overestimator or underestimator) when representing faces of different groups. Specifically, this technique allowed us to explore relationships between trait ratings of GROUP X NEST group-level classification images (i.e., ingroup-overestimator, outgroup-overestimator, ingroup-underestimator, and outgroup-underestimator) and the linear combinations of trait ratings of GROUP and NEST group-level classification images from Study 1a. My premise was that participants completed the face categorization task from Study 1a with two pieces of information: 1) the category label (overestimator or underestimator) of the targets (NEST) and 2) whether the targets shared their group membership or not (ingroup or outgroup - GROUP). The RSA technique uses similarity matrices to examine the relationship between different representational spaces (e.g., the relationship between how ingroup faces are generally rated and how overestimator faces are

generally rated). Each cell in each similarity matrix is a pairwise similarity (e.g., correlation) between two traits (e.g., trustworthy and attractive). Quantitatively, if participants had only those two pieces of information (GROUP and NEST) at hand during the face categorization task and indeed used them, trait representational space of GROUP X NEST images should reflect linear combinations of trait representations of GROUP images and those of NEST images. Thus, by using multiple regression RSA, I attempted to tease apart unique contributions of category labels (NEST) and whether the target shared the same group membership with the participant or not (GROUP) in how people chose faces who belonged to one of four GROUP X NEST groups (ingroup-overestimator, outgroup-overestimator, ingroup-underestimator, and outgroup-underestimator) during the face categorization task.

Methods

Participants. The data from the same 301 participants recruited in Phase 2 of Analysis 1 were reanalyzed here. As stated in above, 99 participants rated ingroup and outgroup classification images, and 102 participants rated overestimator and underestimator classification images. Additionally, 100 participants rated ingroup-overestimator, outgroup-overestimator, ingroup-underestimator, and outgroup-underestimator classification images. See the Participants section of Phase 2 of Analysis 1 for a more detailed description.

Procedure. To quantitatively examine contributions of GROUP and NEST in the face categorization task, I first computed pairwise correlations of trait rating data from Study 1a (e.g., correlation between trustworthy and attractive ratings), generating a correlation matrix for each group-level classification image. I then vectorized unique pairwise correlation matrices (i.e., excluding duplicate correlation coefficients). Finally, I used multiple regression RSA to predict correlation vectors of GROUP X NEST trait rating data

with linear combinations of correlation vectors of appropriate GROUP and NEST trait rating data. For example, I predicted trait representations (i.e., a vector of unique pairwise correlation coefficients) of the ingroup-overestimator group-level classification image using the linear combination of trait representations of the ingroup group-level classification image and that of the overestimator group-level classification image (see Figure 4). By using multiple regression RSA, I tested unique contributions of group membership (GROUP) and minimal group labels (NEST) in GROUP X NEST images while controlling for each other. If participants used one type of information more than the other, it should yield a higher slope value. For example, if people used the ingroup/outgroup distinction more than the overestimator/underestimator distinction when choosing ingroup underestimator faces during the face categorization task, the trait representations of ingroup should have a higher beta value than the trait representation of underestimator. All de-identified data, analysis scripts, and study materials are posted at https://osf.io/s9243.

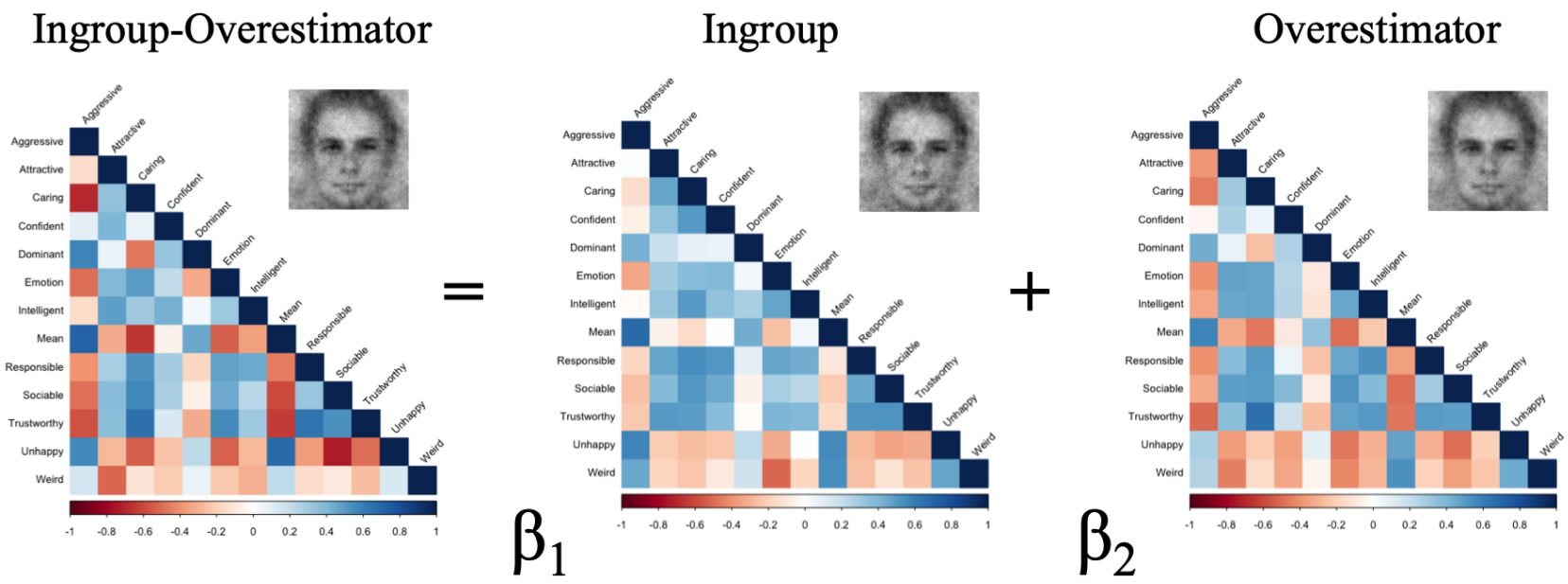


Figure 4. Multiple regression RSA example: predicting the pairwise correlation matrix of ingroup overestimator trait rating data from the linear combination of the pairwise correlation matrices of ingroup trait rating data and overestimator trait rating data. Each square represents a pairwise correlation value.

Results

Ingroup overestimator. I used ordinary least squares multiple regression to predict unique pairwise correlation vectors of thirteen trait rating data of ingroup overestimator face image with the linear combination of the correlation vectors of ingroup face trait rating data and correlation vectors of overestimator face image trait rating data. I found that both the ingroup ratings ($\beta = .206$, $SE = .118$, $t(77) = 2.279$, $p = .026$) and overestimator ratings ($\beta = .752$, $SE = .099$, $t(77) = 8.303$, $p < .001$) were significant predictors of ingroup overestimator ratings. I also conducted linear hypothesis testing to test whether ingroup ratings and overestimator ratings were significantly different from each other and found that overestimator ratings predicted ingroup overestimator ratings significantly better than ingroup ratings ($F(1,75) = 6.811$, $p = .011$).

Outgroup overestimator. I followed the same procedures described above for the ingroup overestimator to predict the outgroup overestimator trait rating data with the linear combination of the correlation vectors of outgroup face trait rating data and correlation vectors of overestimator face image trait rating data. I found that both the outgroup ratings ($\beta = .172$, $SE = .081$, $t(77) = 3.475$, $p < .001$) and overestimator ratings ($\beta = .824$, $SE = .055$, $t(77) = 16.632$, $p < .001$) were significant predictors of outgroup overestimator ratings. Linear hypothesis testing showed that overestimator ratings predicted outgroup overestimator ratings significantly better than outgroup ratings ($F(1,75) = 23.696$, $p < .001$).

Ingroup Underestimator. I used multiple regression to predict ingroup underestimator trait rating data with the linear combination of ingroup face trait rating data and underestimator face image trait rating data. I found that both the ingroup ratings ($\beta = .534$, $SE = .058$, $t(77) = 10.131$, $p < .001$) and underestimator ratings ($\beta = .497$, $SE = .054$, $t(77) =$

9.418, $p < .001$) were significant predictors of ingroup underestimator ratings. The linear hypothesis testing showed that ingroup ratings and underestimator ratings did not significantly differ in predicting ingroup underestimator ratings ($F(1,75) = .574$, $p = .451$).

Outgroup Underestimator. I used multiple regression to predict outgroup underestimator trait rating data with the linear combination of outgroup face trait rating data and underestimator face image trait rating data. I found that both the outgroup ratings ($\beta = .496$, $SE = .078$, $t(77) = 7.084$, $p < .001$) and underestimator ratings ($\beta = .482$, $SE = .073$, $t(77) = 6.887$, $p < .001$) were significant predictors of outgroup underestimator ratings. Linear hypothesis testing showed that outgroup ratings and underestimator ratings did not significantly differ in predicting outgroup underestimator ratings ($F(1,75) = .137$, $p = .712$).

Discussion

Analysis 2 examined how people generated face representations of different minimal groups using multiple regression RSA on trait rating data of group-level classification images. I found that for both ingroup overestimator and outgroup overestimator face representations, participants seemed to have used the overestimator label more than the ingroup/outgroup dimension, as indicated by larger trait representational similarities between GROUP X NEST face images (i.e., ingroup-overestimator and outgroup-overestimator) and the overestimator face image than the ingroup or outgroup face image (i.e., larger β values for overestimator trait representations than ingroup or outgroup trait representations). On the other hand, for ingroup underestimator and outgroup underestimator face representations, participants seemed to have used group membership (ingroup or outgroup) and the underestimator label equally, as indicated by equally similar trait representations between

GROUP X NEST face images (i.e., ingroup-underestimator and outgroup-underestimator) and the underestimator face image and ingroup or outgroup face image.

The larger role that the overestimator label played during the face categorization task can also be interpreted, both conceptually and mathematically, that trait representations of ingroup overestimator and outgroup overestimator faces were similar. In other words, ingroup overestimator and outgroup overestimator face representations elicited overall similar trait impressions from an independent sample of participants. In contrast, ingroup and outgroup underestimator face representations did not show as much correspondence in their trait representations with each other. This may suggest that in the aggregate, people have more consistent representations of overestimator faces (i.e., consensus across participants) compared to underestimator faces. Together these findings showed that minimal group labels may indeed be meaningful when visualizing faces, but different labels may have different levels of influence.

It is important to note that my interpretations of multiple regression RSA results are drawn from trait ratings of group-level classification images. Although past research suggests that trait impressions and behaviors elicited by group-level classification images resemble those elicited by participant-level classification images (Dotsch et al., 2008; Ratner et al., 2014), recent studies show that examining differences between group-level classifications only might be susceptible to inflated Type I error rate (Cone et al., 2020). Thus, I examined whether the difference I found in mental representations of overestimator and underestimator faces holds at the participant level in Analysis 3.

Analysis 3

Analysis 1 and 2 provided evidence that face representations of minimally defined groups can vary and lead to trait impressions that differ on various dimensions of face perception and that different minimal group labels have different degree of influence on people's mental representation of ingroup and outgroup faces. One potential limitation to these findings is that we assessed trait impressions of group-level classification images, which are the summary representation (i.e., average) of many participant-level classification images. Although this summary representation of what the face of a given group member (e.g., overestimator) might very well represent most of the cases that make up the average, using summary representations does not necessarily indicate that the individual participants were actually visualizing ingroup and outgroup members differently as a function of the specific group labels.

In Analysis 3, I tested whether representational differences found in trait impressions of different minimal groups in Study 1a also exist at an individual level by examining the representational differences in participant-level classification images of ingroup and outgroup as well as overestimator and underestimator faces. To do so, I used a machine learning analytic approach, that examines the relationship between pixel intensity data of each image and its category labels, thus circumventing biases that might arise from subjective trait ratings. This approach has not been used previously to examine biases in reverse correlation classification images and is vastly different from the trait impression analytic approach used in Study 1a and 1b. Finding similar representational differences between different categories using this approach would therefore provide strong convergent evidence that the previous Study 1 effects we report are robust.

Methods

Stimuli. In Analysis 3, I used 362 participant-level classification images from Phase 1 of Analysis 1. Each image had three dimensions: (1) GROUP (ingroup or outgroup), (2) NEST (overestimator or underestimator), and (3) GROUP X NEST (ingroup-overestimator, outgroup-overestimator, ingroup-underestimator, or outgroup-underestimator).

Procedure. I used the R package *e1071* (Meyer et al., 2019) to conduct the machine learning analyses following three steps: 1) vectorizing and down sampling pixel intensity data of each image (see Figure 5), 2) standard scaling (i.e., standardization), and 3) classification using support vector machines (SVM) with a radial basis function (RBF) kernel (Cortes & Vapnik, 1995; Scholkopf et al., 1997). I performed cross-validation for each analysis to ensure that every image was in (not at the same time) both training and testing datasets. I down sampled participant-level classification images by simply re-sizing them from 512 x 512 to 64 x 64. We conducted the same true label analysis using 512 x 512, 256 x 256, 128 x 128, and 64 x 64 image sizes, and found no detrimental effect of down sampling on the classification accuracies. Thus, I downsampled the images due to the computationally intensive nature of machine learning analyses needed for the 1000 permutation tests. Standard scaling was done by mean-centering pixel intensity data and dividing the values by their standard deviation. Finally, I used the SVM to classify each image to the appropriate category, using a radial basis function with default cost and gamma hyperparameters (cost = 1 and gamma = 1/n features = 64 x 64). I used this same procedure to classify between ingroup and outgroup faces, overestimator and underestimator faces, and GROUP x NEST faces.

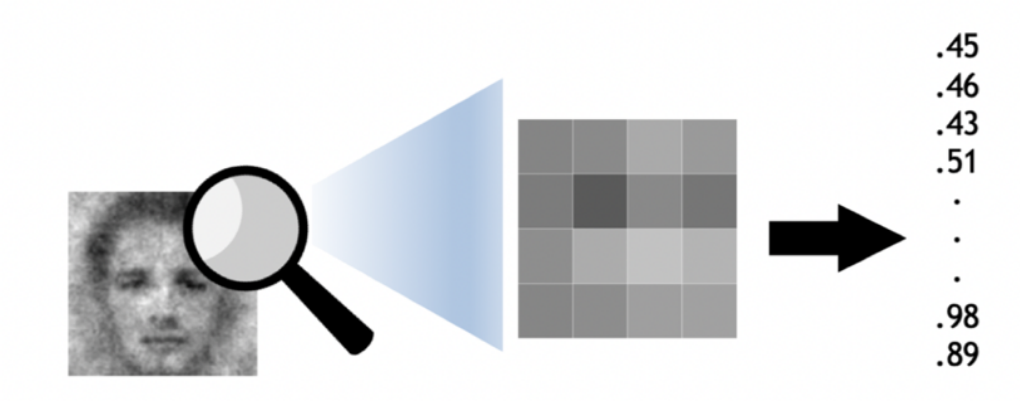


Figure 5. Example of vectorizing pixel intensity data from a participant-level classification image

To minimize overfitting and maximize the chance of detecting real differences in classification images, I used 10-fold cross-validations with my SVM models. Each fold yielded a training set (90% of data = 325 to 326 cases) and a testing set (10% of data = 36 to 37 cases), both of which were evenly divided between classes (e.g., approximately equal numbers of ingroup and outgroup images). The SVM algorithm then learned the relationships between features (64 x 64 vectorized pixel intensities of each image) and class labels (e.g., ingroup and outgroup) from the training set, and classified images from the testing set consisting of images that were not part of the training set for a given fold. I repeated this step 10 times until every instance of data was in both training and testing sets at some point. I then computed accuracy scores by averaging accuracies from these 10 folds. By using this method of cross-validation, I ensured that class labels were balanced for both training and testing sets, and no image was included in both training and testing sets at the same time for any given fold.

Next, I used permutation tests to determine whether accuracies of my SVM classification results differed significantly from chance (Ojala & Garriga, 2009). For each permutation, class labels (e.g., ingroup or outgroup) were randomly permuted for every

image, followed by the classification steps described above. I repeated the same procedure 1000 times, creating my own null distribution against which I could compare the accuracies of my classification results with true labels. I then estimated the p-value from the proportion of permutation accuracies that exceeded the accuracy with true labels (i.e., percentage of permutation tests that had higher accuracy than the accuracy with true labels).

I then compared the accuracies of my model's classifications of GROUP and NEST labels using the 5 X 2-fold cross-validation paired samples t-test (Dietterich, 1998). That is, I performed five replications of 2-fold cross-validations (splitting data into equal number of training and testing data), resulting in ten accuracy scores for each classification. We then used a simple paired samples t-test on those accuracy scores to test whether the model performed significantly better classifying GROUP or NEST labels. I did not use the paired samples t test on accuracy scores from the 10-fold cross-validation because it violates a key assumption of the t-test. Specifically, for 10-fold cross-validation, an instance of data is used in the training set 9 times, and therefore accuracy scores are not independent from each other. This in turn leads to inflation of Type I error (Dietterich, 1998). With the 5 X 2-fold cross-validation, each instance of data appears only in the training or testing set for any given fold, ensuring independence between accuracy scores, thus reducing the likelihood of Type I error. All de-identified data, analysis scripts, and study materials are posted at https://osf.io/s9243/.

Results

I was able to classify between ingroup and outgroup images from pixel intensity data significantly better than chance (accuracy = 59.20%, $p < 0.001$). The same was true for overestimator and underestimator images (accuracy = 66.01%, $p < 0.001$). For both classifications, all permutation tests yielded lower accuracy scores than the accuracy scores

with true labels. Next, I compared classification accuracies for GROUP and NEST using the 5 X 2-fold cross-validation paired samples t-test. This resulted in slightly different accuracy scores for each classification from 10-fold cross-validation accuracy scores (GROUP = 55.75% and NEST = 62.88%). The t-test result showed that the model performed significantly better classifying NEST labels than GROUP labels, $t(9) = 4.65$, $p = .001$, two-tailed, Cohen's $d = 1.47$, 95% CI of difference [3.66, 10.60].

Finally, multiclass SVM results showed that the model performed significantly better than chance (accuracy = 37.83 %, $p < 0.001$). Unlike the previous two cases, the chance accuracy for the current classification was 25% (1 out of 4). Upon examining the confusion matrix of results using the true labels, I **found** that my model misclassified within NEST labels (99/225) more than within GROUP labels (72/225), such as classifying ingroup overestimator face images as outgroup overestimator rather than ingroup underestimator.

Discussion

In Analysis 3, I investigated whether the difference between face representations of overestimators and underestimators exists not just in the aggregate trait-rating data but also at an individual level in the pixel intensity data. I examined the differences between GROUP (ingroup and outgroup), NEST (overestimator and underestimator), and GROUP X NEST participant-level classification images using a novel approach for analyzing reverse correlation images, specifically a machine learning algorithm called support vector machine. I found that using this method I could classify all GROUP, NEST, and GROUP X NEST participant-level classification images better than chance, suggesting that the differences between face representations of all category types exist at the individual level.

I also found that the SVM classified between overestimator and underestimator face images significantly better than ingroup and outgroup face images, providing evidence that NEST labels were used more than ingroup-outgroup status during the face categorization task, resulting in more consistent face representations of overestimator and underestimator than those of ingroup and outgroup across different participants. One explanation of this effect is that the face categorization task had an explicit task goal of choosing overestimator or underestimator faces, whereas group membership was implicit - whether the participant shares the same group membership with the targets or not. Thus, this might have contributed to more consistent face representations of overestimator and underestimator than ingroup and outgroup. However, the results of Analysis 2 may partially address this possibility: the so-called more "consistent" face representation of overestimator and underestimator than ingroup and outgroup found in Analysis 3 was true mostly for overestimator faces but not necessarily for underestimator faces in Analysis 2. Thus, we argue that category labels can be meaningful, albeit to different extents for different labels (i.e., overestimator is more meaningful than underestimator).

In short, I showed that the representational differences between ingroup and outgroup as well as overestimator and underestimator exist not only in the summary representations (i.e., average of many participant-level classification images), but also in individuals' face representations of different groups. I also did not use subjective trait ratings to arrive at this conclusion, thus providing converging evidence that ingroup and outgroup faces as well as overestimator and underestimator faces are objectively different from each other. I was also able to show the same findings as Analysis 1 despite using two very different methods (i.e., trait ratings vs. image classification using pixel intensity data), suggesting that the findings of

top-down influences of ingroup positivity and category labels on face representations are robust.

The finding of more misclassifications within NEST labels than within GROUP labels of GROUP X NEST participant-level classification images provided another piece of evidence that people might have used NEST labels more than group membership when visualizing faces during the face categorization task. Although these findings are descriptive, the category labels seemed to have played greater roles in the face categorization task than whether the targets shared the same group membership with the participant or not.

Study 1b

So far, I showed that people have different mental representations for different minimal groups, and that this difference may be driven more by people's mental representations of what an overestimator should look like rather than what an underestimator should look. Thus, people seem to imbue meaning to minimal group labels but to different extents for different labels. One critical limitation to my findings is that I used only one version of the minimal group paradigm (i.e., the overestimator versus underestimator distinction), therefore I have not shown whether people would imbue meaning to other minimal group labels (e.g. Klee versus Kandinsky fans). Additionally, although I showed that one type of minimal group paradigm can be meaningful to some people, it is still unclear what the implications of that are for research using the minimal group paradigm to investigate various forms of intergroup bias. In the minimal group literature, all versions of the minimal group paradigm are typically viewed as different means to the same end. That is, they have their own unique ways of manipulating novel group memberships, but they are interchangeable and whether one version or another is used is often left to the preferences of

the researchers. However, if you take the possibility seriously that perceivers might be motivated to read into category labels, then this raises the question of whether some minimal group operationalizations have more inductive potential than others. For instance, as shown above, it is rather clear to see how overestimators and underestimators are viewed differently. However, what about people who prefer paintings by Klee versus Kandinsky? It is easy to see how based on associations between ethnicity and surname that Klee might be assumed to be Western European, and Kandinsky might be assumed to be Eastern European. All the stereotypes associated with these groups could then become accessible. However, it does not logically follow that people who prefer one abstract artist or the other would share their preferred artist's ethnicity. Thus, unlike in the case of overestimators and underestimators, it is hard to overtly reason how people who prefer one abstract artist versus another differ from each other. For this reason, a close look at the overestimator/underestimator and Klee/Kandinsky paradigms illustrates how category labels across minimal group paradigms have different inductive potential. By inductive potential, I mean the ease at which people can infer meaning from the labels. This is not different from the reality that some groups in the real world (even when these groups are otherwise novel) have names and other attributes that allow characteristics about group members to be inferred more easily than do names of other groups.

It is unclear what the implications of varying degrees of inductive potential in different minimal group paradigms would be for various intergroup responses. On the one hand, based on previous findings showing that more arbitrary group distinctions led to little to no discrimination between ingroup and outgroup (e.g., Rabbie & Horwitz, 1969; Billig & Tajfel, 1973), we can expect to see less discrimination between ingroup and outgroup when

the inductive potential of the labels is low as is the case for the Klee and Kandinsky groups based. On the other hand, it is not unreasonable to expect the opposite pattern of results. When the two labels (e.g., Klee and Kandinsky) are less distinguishable, the minimal group situation becomes more ambiguous. One piece of clear information to the participants is that whether the target shares the same group membership with them (ingroup) or not (outgroup). Thus, people may focus more on the ingroup/outgroup distinction rather than reading into the category labels, thus leading to greater discrimination between ingroup and outgroup.

Thus, Study 1b attempted to replicate the findings from Study 1a with a different type of minimal group paradigm, the Klee versus Kandinsky distinction (Experiment 2 from Tajfel et al., 1971). Study 1b used the same set of methods from Study 1a to empirically examine whether people represent faces of people who like Klee paintings differently from faces of those who like Kandinsky paintings. Following the procedure of Study 1a, Study 1b was also conducted in three parts (Analysis 1,2, and 3).

Analysis 1

Participants were first randomly assigned to minimal groups (Klee fans versus Kandinsky fans) and then categorized faces as belonging to either of these two minimal groups. I again used the reverse correlation image classification technique to create visual representation of Klee and Kandinsky fans as well as ingroup and outgroup faces. For Analysis 1, I assessed whether different images of these different minimal group faces would be rated differently by independent samples of participants on the thirteen trait dimensions (Oosterhof & Todorov, 2008).

Although I found some differences in trait impressions between different minimal groups (overestimator versus underestimator) from Study 1a, I chose to remain agnostic

about whether the Klee and Kandinsky group labels would result in different face representations because it is possible that these labels used for the aesthetic preference minimal group paradigm have less inductive potential. However, given that this version of the minimal group paradigm has revealed ingroup favoritism in past research (e.g., Tajfel et al., 1971), I still predicted that people would show ingroup positivity as indicated by more positive trait ratings of ingroup faces than outgroup faces.

Method

Phase 1: Generating visual renderings of face representations

Participants. I recruited 200 University of California Santa Barbara students (M_{age} = 18.82, SD = 1.07; 149 female, 47 male, and 4 unidentified) to participate in a study about categorizing faces in exchange for course credit via the UCSB Psychological and Brain Sciences subject pool. Racial and ethnic breakdown of participants was 65 Asian, 65 White, 35 Latinx, 24 multiracial, 5 other, and 6 unidentified. Up to four participants were run simultaneously.

Procedure. The current study followed the same exact procedure as Study 1a except for the version of the minimal group paradigm used to assign participants to different groups. As with Study 1a, participants were first told that they would perform several tasks on a computer. Next, I used a classic *aesthetic preference* procedure (Experiment 2 from Tajfel et al., 1971) to assign participants to arbitrary, but believable, groups. Then they conducted a face categorization task optimized for a reverse correlation analysis.

Artistic Preference Test (ART). In this task, I told the participants that people can reliably figure out another person's artistic preference simply by looking at their face. I then told the participants that they would categorize photographs of students from a previous

quarter whose artistic preference had been determined. I also told them that the purpose of the current study was to test whether people can determine artistic preference when faces appear blurry.

Next, participants completed the artistic preference test themselves. In this task, they viewed 12 pairs of paintings (a pair per trial) by modern European artists, Paul Klee and Wassily Kandinsky, and chose whichever painting they liked better on a given trial. On each trial, one of the paintings was by Kandinsky and the other one was by Klee. The location of each painting (whether on the left or right of the screen) did not correspond to the painter, and the signature of the painter was hidden from each painting to prevent participants from choosing on the basis of the painter's name. At the end of the test, the computer program provided predetermined feedback (counter-balanced across participants), indicating that each participant had a preference for paintings by either Kandinsky or Klee. As was the case in the numerical estimation style test from Study 1a, I did not actually take participants responses into account; it was used to provide a rationale for the group assignment.

I used additional procedures to make the novel group membership (i.e., Klee and Kandinsky) as salient as possible in participants' minds throughout the remainder of the study. First, participants reported their artistic preference to the experimenter, and the experimenter then wrote each participant's identification number and artistic preference on a post-it note and attached it to the bottom center of the computer monitor (in the participants' line of sight) to constantly remind them of their group membership during the face categorization task. Participants also typed their artistic preference into the computer.

Face categorization. After the group assignment, participants completed a forced-choice face categorization task for 450 trials. On each trial, participants selected a face of

someone who prefers paintings by either Kandinsky or Klee out of two adjacent grayscale face images. Half of the participants were asked on every trial to choose which of the two faces belonged to a person who preferred Kandinsky and the other half of the participants were asked on every trial to choose the person who preferred Klee. If the targets shared the same artistic preference as the participant, then the participant was selecting ingroup faces, whereas if the targets did not share the artistic preference with the participant, then the participant was selecting outgroup faces. I used the same set of face stimuli from Study 1a – 450 pairs of face images generated from grayscale neutral male average face of the Averaged Karolinska Directed Emotional Faces Database (Lundqvist et al., 1998). I presented inverse noise faces equally on the left and right sides of the screen in a random order. I used the same pairs of faces for all participants. The rest of the procedure, including the individual differences questionnaire and taking photographs of the participants with their consent remained the same in this study. For the individual difference questionnaires, both the PSE and CSE yielded reasonable Cronbach's α levels (.91 and .85).

Face representation data processing. I used the same reverse correlation analysis from Study 1a to generate visual renderings of different groups by averaging noise patterns of selected faces – Klee and Kandinsky group faces, ingroup and outgroup faces, and Klee-ingroup, Klee-outgroup, Kandinsky-ingroup, and Kandinsky-outgroup faces. I generated both participant-level classification images and group-level classification images (refer back to Methods section of Analysis 1 of Study 1a for a more detailed description of this procedure). To test whether the Klee and Kandinsky version showed the ingroup positivity effect found in Study 1a, I created ingroup (n = 100) and outgroup (n = 100) classification images (see Figure 6). Second, to examine the differences between Klee and Kandinsky groups, I created

Klee (n = 100) and Kandinsky (n = 100) classification images collapsed across ingroup and outgroup (see Figure 7). Finally, I examined the interaction between ingroup/outgroup and Klee/Kandinsky distinctions by creating four classification images by crossing the two dimensions: ingroup-Klee (n = 50), ingroup-Kandinsky (n = 50), outgroup-Klee (n = 50), and outgroup-Kandinsky (n = 50). All four classification images can be seen in Figure 8.

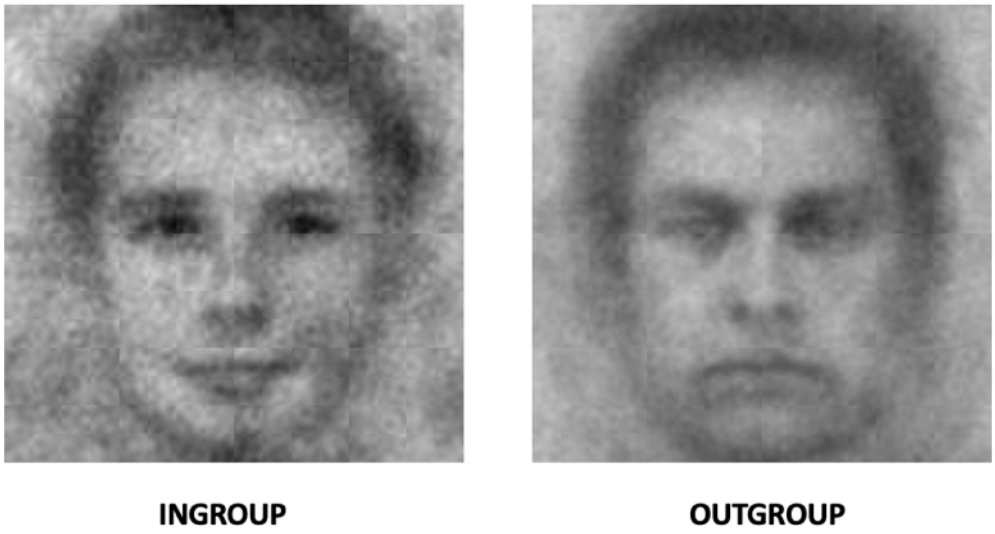


Figure 6. Study 1b ingroup and outgroup group-level classification images

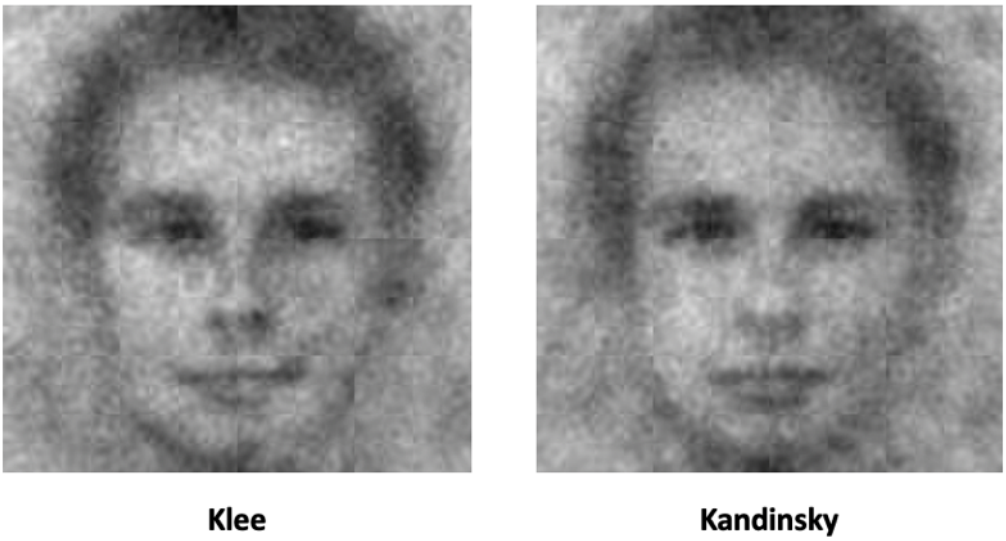


Figure 7. Study 1b Klee and Kandinsky group-level classification images

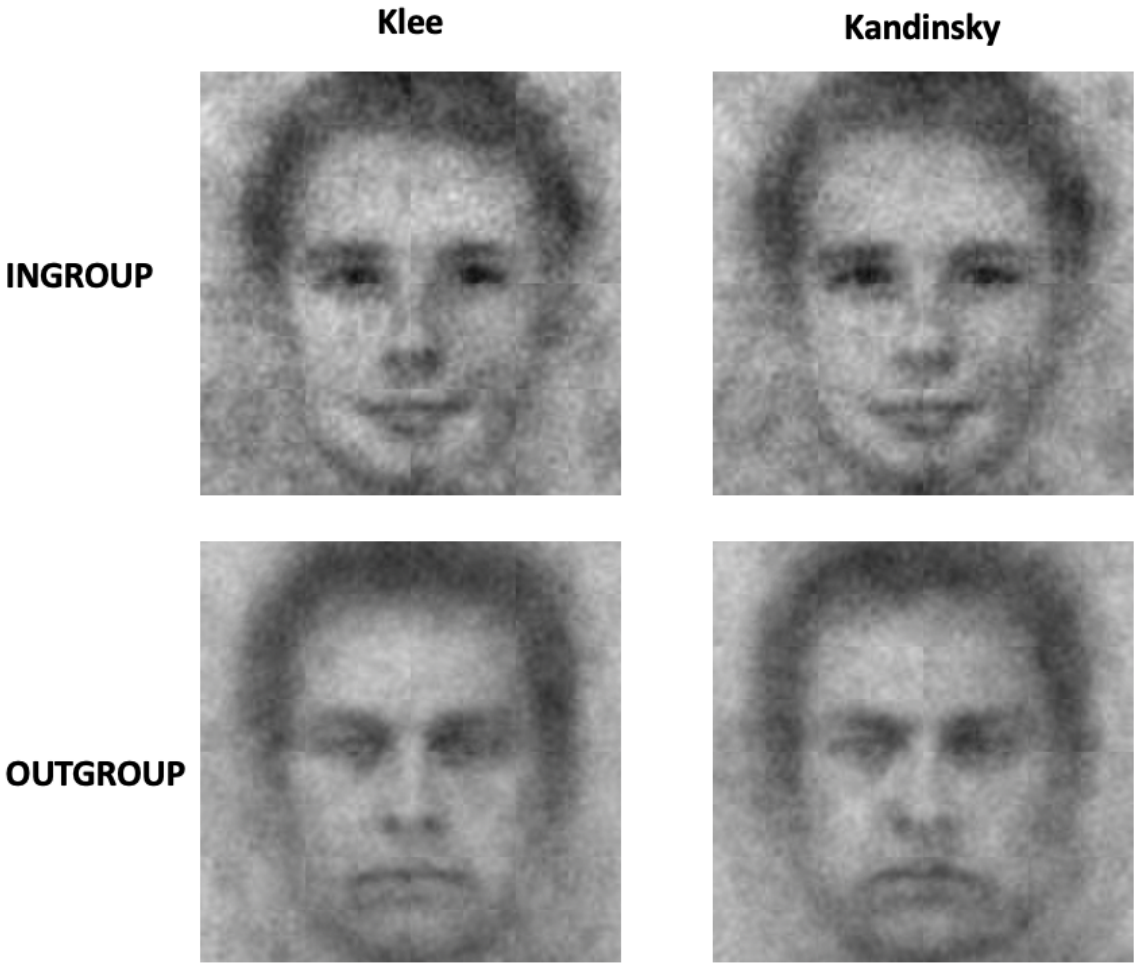


Figure 8. Study 1b GROUP X ART group-level classification images

Phase 2: Assessing impressions of face representations

In Phase 2, I assessed how different face images elicited different trait impressions. Independent samples of participants who were not aware of the face generation phase from Phase 1 rated eight group-level classification images. To assess relative differences between ingroup and outgroup (GROUP), Klee and Kandinsky (ART), and GROUP X ART images, I obtained ratings from three different samples of participants. That is, participants only rated ingroup and outgroup images, Klee and Kandinsky images, or GROUP X ART images.

Participants. I recruited a total of 150 participants (M_{age} = 35.08, SD = 11.15; 96 female, 54 male) through the TurkPrime website (www.turkprime.com) to complete an

online survey administered through Qualtrics (www.qualtrics.com). 50 participants rated ingroup and outgroup classification images, 50 participants rated Klee and Kandinsky classification images, and 50 participants rated ingroup-Klee, outgroup-Klee, ingroup-Kandinsky, and outgroup-Kandinsky classification images. Racial and ethnic breakdown of the raters was 103 White, 26 Black, 6 Latinx, 5 Asian, 1 Native American, 1 Pacific Islander/Hawaiian, and 8 multiracial participants. Participants were expected to complete the study in 10 minutes. All participants did not know about the face categorization stage of the study. They were compensated with $1 for their participation.

Procedure. After providing informed consent, participants rated the classification images on thirteen trait dimensions (i.e., To what extent is this face… trustworthy, attractive, dominant, caring, sociable, confident, emotionally stable, responsible, intelligent, aggressive, mean, weird, and unhappy?) (Oosterhof & Todorov, 2008). Each face was presented by itself in a random order. Ratings were made on scales from 1 (not at all) to 7 (extremely). The order of each trait presentation was also random.

Results

For each sample of raters, I conducted a repeated-measures multivariate analysis of variance (rMANOVA) followed by univariate analysis of variance for each trait. I show the results below separated by sample.

Group membership (GROUP). A rMANOVA comparing the trait ratings of ingroup and outgroup classification images was significant, Pillai's Trace = .83, $F = 14.05$, $df = (13, 37)$, $p < .0001$, indicating some difference in trait ratings between ingroup and outgroup classification images. The univariate F tests showed that all trait ratings of ingroup and outgroup images were significantly different from each other at the .05 significance level.

The means, F values, p values, and effect sizes for each comparison are presented in Table 4. The ingroup face was rated significantly more trustworthy, attractive, caring, emotionally stable, responsible, intelligent, and sociable; the outgroup face was rated significantly more dominant, aggressive, mean, weird, and unhappy.

	Ingroup mean (SD)	Outgroup mean (SD)	F-value	Cohen's d
Trustworthy	5.32 (1.15)	2.58 (1.50)	103.32***	1.44
Attractive	4.74 (1.29)	2.66 (1.48)	59.66***	1.09
Dominant	3.20 (1.50)	5.72 (1.29)	66.92***	1.16
Caring	5.34 (1.42)	2.32 (1.46)	110.08***	1.48
Confident	5.12 (1.42)	3.46 (1.47)	27.76***	.75
Emotionally stable	5.24 (1.20)	2.68 (1.35)	88.07***	1.33
Responsible	5.00 (1.28)	3.58 (1.34)	30.46***	.78
Intelligent	5.00 (1.18)	3.52 (1.40)	40.51***	.90
Aggressive	2.14 (1.47)	5.94 (1.15)	170.09***	1.84
Mean	2.06 (1.46)	5.86 (1.37)	138.20***	1.66
Weird	3.06 (1.85)	3.84 (1.78)	5.86*	.34
Unhappy	2.10 (1.63)	5.98 (1.38)	114.09***	1.51
Sociable	5.78 (1.04)	2.30 (1.61)	145.10***	1.70

Significance codes: *** <.001 ** <.01 * <.05 $^{+}$<.10

Table 4. Study 1b trait rating ANOVA results – GROUP (ART) – face representations

Artistic preference (ART). A rMANOVA comparing the trait ratings of Klee and Kandinsky classification images was significant, Pillai's Trace = .60, $F = 4.32$, df = (13, 37), $p < .001$, indicating some difference in trait ratings between Klee and Kandinsky classification images. The univariate *F* tests showed that the majority of trait ratings of Klee and Kandinsky images were significantly different from each other at the .05 significance level. The means, *F* values, p values, and effect sizes for each comparison are presented in

Table 5. The Klee group face was rated significantly more caring, confident, emotionally stable, and sociable; the Kandinsky group face was rated significantly more aggressive, mean, and unhappy. Trustworthy, attractive, dominant, responsible, intelligent, and weird were not significantly different between Klee and Kandinsky face images at the .05 significance level.

	Klee mean (SD)	Kandinsky mean (SD)	F-value	Cohen's d
Trustworthy	4.82 (1.21)	4.62 (1.23)	1.04	.14
Attractive	4.52 (1.53)	4.12 (1.44)	3.84^{+}	.28
Dominant	4.2 (1.64)	4.14 (1.62)	.10	.05
Caring	4.96 (1.11)	4.22 (1.42)	12.70***	.50
Confident	5.30 (1.16)	4.30 (1.43)	15.91***	.56
Emotionally stable	5.20 (1.34)	4.58 (1.37)	8.74**	.42
Responsible	4.72 (1.29)	4.72 (1.29)	.00	.00
Intelligent	4.88 (1.22)	4.56 (1.18)	2.54	.23
Aggressive	3.08 (1.97)	3.60 (1.80)	4.34*	.29
Mean	3.14 (1.99)	3.94 (1.73)	15.37***	.55
Weird	3.78 (1.96)	3.32 (1.79)	3.12^{+}	.25
Unhappy	2.90 (1.88)	4.54 (1.31)	32.70***	.81
Sociable	5.28 (1.31)	4.20 (1.59)	15.07***	.55

Significance codes: *** <.001 ** <.01 * <.05 $^{+}$<.10

Table 5. Study 1b trait rating ANOVA results – ART – face representations

GROUP X ART. I used rMANOVA to test the effects of GROUP, ART, and the interaction between the two on trait ratings. A significant multivariate effect was found only for GROUP (Pillai's Trace = .66, F = 20.57, df = (13, 135), p < .001). The effects of ART (Pillai's Trace = .09, F = 1.04, df = (13, 135), p = .41) and GROUP X ART (Pillai's Trace = .07, F = .75, df = (13, 135), p = .71) were not significant. Ingroup faces were rated more

trustworthy, attractive, caring, confident, emotionally stable, responsible, intelligent, and sociable, whereas outgroup faces were rated more dominant, aggressive, mean, weird, and unhappy for both Klee and Kandinsky face images. The differences found between the Klee and Kandinsky groups dissipated when the category labels were crossed with group membership. The univariate F test results including the means, standard deviations, F values, p values, and effect sizes (comparing ingroup and outgroup within Klee and Kandinsky) for each trait are presented in Table 6.

	Klee			**Kandinsky**			**F-values**		
	Ingroup (SD)	Outgroup (SD)	Cohen's d	Ingroup (SD)	Outgroup (SD)	Cohen's d	Group	ART	Group X ART
Trustworthy	5.04 (1.48)	3.42 (1.53)	.80	5.22 (1.43)	3.30 (1.74)	.91	97.19***	.03	.70
Attractive	4.62 (1.46)	3.20 (1.67)	.81	4.68 (1.41)	2.72 (1.71)	1.03	112.50***	1.74	2.87
Dominant	3.36 (1.97)	5.28 (1.26)	.84	3.42 (1.72)	5.12 (1.47)	.72	81.43***	.06	.30
Caring	5.34 (1.30)	2.84 (1.73)	1.19	5.38 (1.21)	2.98 (1.88)	.98	157.23***	.21	.07
Confident	5.18 (1.06)	4.16 (1.60)	.56	4.90 (1.40)	3.86 (1.54)	.56	33.61***	2.66	.00
Emotionally stable	5.22 (1.30)	3.52 (1.50)	.98	4.98 (1.50)	3.28 (1.60)	.85	104.83***	2.09	.00
Responsible	4.76 (1.35)	3.92 (1.51)	.46	4.86 (1.23)	3.70 (1.45)	.61	36.11***	.13	.92
Intelligent	4.82 (1.16)	3.84 (1.38)	.73	4.70 (1.37)	3.46 (1.43)	.68	59.12***	3.00	.81
Aggressive	2.68 (1.90)	5.44 (1.05)	1.21	2.56 (1.92)	5.42 (1.31)	1.28	209.46***	.13	.07
Mean	2.68 (1.82)	5.24 (1.20)	1.17	2.70 (1.91)	5.62 (1.23)	1.15	173.52***	.92	.75
Weird	3.34 (1.94)	4.26 (1.75)	.47	3.14 (1.86)	4.50 (1.75)	.60	34.60***	.01	1.29
Unhappy	2.72 (1.97)	5.66 (1.39)	1.24	2.88 (1.81)	5.96 (1.34)	1.24	192.54***	1.12	.10
Sociable	5.46 (1.33)	2.90 (1.63)	1.27	5.26 (1.29)	2.82 (1.79)	1.10	176.45***	.55	.10

Significance codes: *** <.001 ** <.01 * <.05 $^{+}$<.10

Table 6. Study 1b trait rating ANOVA results – GROUP X ART – face representations

Discussion

In Analysis 1 of Study 1b, I investigated the generalizability of the findings from Analysis 1 of Study 1a to a different type of minimal group paradigm. First, I replicated the ingroup positivity effect in face representations: ingroup faces elicited overall more positive trait impressions compared to outgroup faces. It is also notable that the magnitude of ingroup positivity was greater for a majority of the traits in the Klee and Kandinsky version compared to the overestimator and underestimator version. That is, the effect sizes were greater for trustworthy, attractive, dominant, caring, emotionally stable, aggressive, mean, and sociable, indicating that in the Klee and Kandinsky version compared to the overestimator and underestimator version, the ingroup face elicited even more positive trait impressions than the outgroup face. This is despite the fact that the sample sizes were smaller in this study compared to that of Study 1a.

I also found some support for the generalizability of Study 1a category label findings: the Klee group face was rated more caring, confident, emotionally stable, and sociable, whereas the Kandinsky group face was rated more mean and unhappy. I did not expect to find more favorable trait impressions for the Klee group face compared to the Kandinsky group face, thus I am hesitant to interpret why this pattern emerged. Nevertheless, my findings still demonstrate that different category labels are represented differently regardless of whether they are ingroup or outgroup, supporting the idea that people may imbue meaning to minimal groups when they are visually representing faces of ingroup and outgroup members. Interestingly, when I crossed category labels with group membership, the differences between the Klee group and the Kandinsky group disappeared. This may be in part due to strong GROUP effects overshadowing the effects of minimal group labels, but

also because people might have focused *more* on whether the target shared the same group membership with them or not, rather than reading into category labels, which was the case for participants in Study 1a. Regardless, the current findings show that people may infer different traits from category labels in different types of minimal group paradigm, but that the inductive potential afforded by the category labels may moderate the extent to which ingroup positivity biases are expressed (i.e., more inductive potential leads to less ingroup bias).

Analysis 2

In Analysis 2, I used multiple regression RSA to examine unique contributions of minimal group labels (ART) and whether the target shared the same group membership with the participant or not (GROUP) in how participants chose faces that belonged to one of four GROUP X ART groups (i.e., ingroup-Klee, outgroup-Klee, ingroup-Kandinsky, and outgroup-Kandinsky) during the face categorization task in Study 2a. Participants had only two pieces of information to complete the task: 1) target minimal group label of the targets (Klee fan or Kandinsky fan) and 2) whether the targets shared the same group membership with them or not (ingroup or outgroup). Because I did not find any effects of category labels in GROUP X ART face images, I expected to find greater contributions of GROUP than ART in trait representations of GROUP X ART images.

Methods

Participant. The data from the same 150 participants recruited in Phase 2 of Analysis 1 were reanalyzed here. 50 participants rated ingroup and outgroup classification images, 50 participants rated Klee and Kandinsky classification images, and 50 participants rated

ingroup-Klee, outgroup-Klee, ingroup-Kandinsky, and outgroup-Kandinsky classification images. See the Participants section of Phase 2 of Study 1b for a more detailed description.

Procedure. I followed the same steps of multiple regression RSA outlined in Study 1a to examine contributions of GROUP and ART in the face categorization task while controlling for each other: (1) I computed pairwise correlations of trait rating data for each group-level classification image, (2) vectorized unique pairwise correlation matrices (i.e., excluding duplicate correlation coefficients), and (3) predicted vectors of GROUP X ART trait rating data with linear combinations of vectors of corresponding GROUP and ART trait rating data. All de-identified data, analysis scripts, and study materials are posted at https://osf.io/s9243/.

Results

Ingroup Klee. I used ordinary least squares multiple regression to predict unique pairwise correlation vectors of thirteen trait rating data of ingroup Klee face image with the linear combination of the correlation vectors of ingroup face trait rating data and correlation vectors of Klee face image trait rating data. I found that both the ingroup ratings ($\beta = .66$, $SE = .09$, $t = 6.08$, $p < .001$) and Klee ratings ($\beta = .24$, $SE = .13$, $t = 2.15$, $p = .035$) were significant predictors of ingroup Klee face image ratings. I also conducted linear hypothesis testing to test whether ingroup ratings and Klee ratings were significantly different from each other and found that ingroup ratings and Klee ratings did not significantly differ in predicting ingroup Klee ratings ($F(1,75) = 1.50$, $p = .225$).

Outgroup Klee. I used multiple regression to predict outgroup Klee trait rating data with the linear combination of ingroup face trait rating data and Klee face image trait rating data. I found that the outgroup ratings ($\beta = .89$, $SE = .05$, $t = 14.71$, $p < .001$) were a

significant predictor of outgroup Klee ratings, whereas the Klee ratings were not (β = .03, SE = .07, t = .57, p = .570). The linear hypothesis testing showed that outgroup ratings predicted outgroup Klee ratings significantly better than Klee ratings (F (1,75) = 36.76, p < .001).

Ingroup Kandinsky. I used multiple regression to predict ingroup Kandinsky trait rating data with the linear combination of ingroup face trait rating data and Kandinsky face image trait rating data. I found that both the ingroup ratings (β = .83, SE = .04, t = 16.45, p < .001) and Kandinsky ratings (β = .19, SE = .08, t = 3.74, p < .001) were significant predictors of ingroup Kandinsky ratings. The linear hypothesis testing showed that ingroup ratings predicted ingroup Kandinsky ratings significantly better than Kandinsky ratings (F (1,75) = 15.34, p < .001).

Outgroup Kandinsky. I used multiple regression to predict outgroup Kandinsky trait rating data with the linear combination of outgroup face trait rating data and Kandinsky face image trait rating data. I found that the outgroup ratings (β = .88, SE = .06, t = 15.15, p < .001) were a significant predictor and of outgroup Kandinsky ratings, whereas the Kandinsky ratings were not (β = .04, SE = .11, t = .76, p = .45). The linear hypothesis testing showed that outgroup ratings predicted outgroup Kandinsky ratings significantly better than Kandinsky ratings (F (1,75) = 27.40, p < .001).

Discussion

In Analysis 2, I found that for all four GROUP X ART face representations, participants seemed to have used the ingroup/outgroup distinction more than the category labels (Klee and Kandinsky) as indicated by larger trait representation similarities between GROUP X ART images and the ingroup and outgroup face images than the Klee or

Kandinsky face image[3]. Although there is evidence that a distinction was made between Klee and Kandinsky in mental representations of ingroup and outgroup faces at the group-level, the current findings are unlike the findings of Study 1a, in that participants in all conditions seemed to have focused more on whether the targets shared their group or not. Thus, I argue that although category labels can be particularly meaningful when visualizing faces (e.g., the overestimator label in Study 1a), the labels with less inductive potential might reveal less label effects, even during a task (e.g., face visualization) that demands forming a concrete representation of the target.

The same limitations of Study 1a apply to the current study. My interpretations of multiple regression RSA results were drawn from trait rating results of group-level classification images, thus may be prone to human bias and only representative of the most typical (i.e., average) face representations. Additionally, it is not clear why Study 2a showed no interaction between GROUP and ART, but the current study suggested that representations of the labels contribute to representations of ingroup but not outgroup faces (i.e., category labels were significant predictors only for ingroup faces). This may simply be due to strong effects of group membership (ingroup/outgroup distinction) overshadowing the effects of category labels (Klee and Kandinsky), but it is also possible that the minimal group labels in this version of minimal group paradigm are only weakly represented.

Analysis 3

Analysis 1 and 2 showed that there might be some differences between face representations of Klee and Kandinsky groups, but whether the targets were ingroup or

[3] For ingroup-Klee face trait representation, the difference between ingroup trait representation and Klee trait representation was not significant ($p = .225$), but the effect size of ingroup trait representation ($\beta = .66$) was more than two times greater than Klee ($\beta = .24$) trait representation.

outgroup played a bigger role than the minimal group labels, which is in opposition with the findings of Study 1a. Furthermore, I found that the magnitude of differences between ingroup and outgroup was larger in this version of minimal group paradigm, indicating that different degrees of meaningfulness of minimal group labels may moderate intergroup bias in face representations of ingroup and outgroup. Thus, in Analysis 3 I examined the representational differences in participant-level classification images of ingroup and outgroup as well as Klee group and Kandinsky group faces by using the support vector machine classifiers. Unlike the findings of Study 1a, I expected the algorithm to perform better in classifying between ingroup and outgroup faces than Klee and Kandinsky group faces based on larger differences found between ingroup and outgroup trait ratings than Klee and Kandinsky trait ratings.

Methods

Stimuli. I used 200 participant-level classification images from Phase 1 of Analysis 1. Each image had three dimensions: (1) GROUP (ingroup or outgroup), (2) ART (Klee or Kandinsky), and (3) GROUP X ART (ingroup-Klee, outgroup-Klee, ingroup-Kandinsky, or outgroup-Kandinsky).

Procedure. I followed the same steps to conduct the machine learning analyses as described in Study 1a. To recap, I 1) vectorized and down sampled pixel intensity data of each image, 2) standardized the data, and 3) performed classification using support vector machines (SVM) with a radial basis function kernel (with default cost and gamma hyperparameters). I used 10-fold cross-validation for each analysis (i.e., classifying between ingroup and outgroup faces, Klee and Kandinsky faces, and GROUP X ART faces). That is, each fold yielded a training set (90% of data = 180 cases) and a testing set (10% of data = 20 cases), both of which were evenly divided between classes (e.g., approximately equal

numbers of ingroup and outgroup images). The SVM algorithm then learned the relationships between features (pixel intensity data) and class labels (e.g., Klee and Kandinsky) from the training set, and classified images from the testing set. I repeated this step 10 times. I then computed accuracy scores by averaging accuracies from these 10 folds. Next, I tested whether the classifiers performed better than chance by using 1000 permutation tests for each analysis. I also compared the accuracy of classification between ingroup and outgroup faces with that of classification between Klee group and Kandinsky group faces using the 5 X 2-fold cross-validation paired samples t-test (Dietterich, 1998). Please see the Procedure section of Analysis 3 of Study 1a for more details. All de-identified data, analysis scripts, and study materials are posted at https://osf.io/s9243/.

Results

The SVM model classified between ingroup and outgroup images from pixel intensity data significantly better than chance (accuracy = 80.00%, $p < .001$). However, the model failed to classify between Klee and Kandinsky group face images better than chance at a .05 significance level (accuracy = 57.00%, $p = .08$). Next, I compared classification accuracies for GROUP and ART using the 5 X 2-fold cross-validation paired samples t-test. This resulted in slightly different accuracy scores for each classification from 10-fold cross-validation accuracy scores (GROUP = 79.40% and ART = 55.30%). The t-test result showed that the model performed significantly better classifying between ingroup and outgroup faces than between Klee and Kandinsky group faces, $t(9) = 12.76$, $p < .001$, two-tailed, Cohen's $d = 4.04$, 95% CI of difference [19.83, 28.37].

Finally, multiclass SVM results showed that the model performed significantly better than chance (accuracy = 41.00 %, $p < .001$). Unlike the previous two cases, the chance

accuracy for the current classification was 25% (1 out of 4). The confusion matrix of results showed that the model misclassified within GROUP (78/160) more than within ART labels (22/104), such as misclassifying ingroup Klee face images as ingroup Kandinsky rather than outgroup Klee.

Discussion

In Analysis 3, I investigated whether the pattern of results found in Analysis 1 and 2 (e.g., slight differences between Klee and Kandinsky, larger differences between ingroup and outgroup) holds true at the participant level using a machine learning analysis. I was able to classify between ingroup and outgroup participant-level classification images but failed to classify between Klee and Kandinsky group images better than chance (i.e., $p = .08$). Considering both Study 1a and Study 1b together, the fact that differences between face representations of ingroup and outgroup exist across different types of minimal group paradigm at both participant and group levels suggest that the ingroup positivity effect in face representation is a robust phenomenon and unlikely to be paradigm specific.

The finding that SVM classified between ingroup and outgroup face images significantly better than Klee and Kandinsky group face images, is contrary to the findings of Study 1a that the SVM model classified between category labels (i.e., overestimator and underestimator) significantly better than ingroup and outgroup. This discrepancy between the two studies supports the idea that different minimal group labels have different degrees of meaningfulness (e.g., NEST labels are more meaningful than ART labels) and that when the minimal group labels have less meaning, people may differentiate between ingroup and outgroup more. This interpretation is also consistent with the finding of more misclassifications within GROUP labels than within ART labels of GROUP X ART

participant-level classification images. Although descriptive, it suggests that people might have used the ingroup/outgroup distinction more than the minimal group labels when visualizing faces during the face categorization task, leading to more consistent face representations of ingroup and outgroup than Klee and Kandinsky groups (e.g., ingroup Klee and ingroup Kandinsky are more similar than ingroup Klee and outgroup Klee).

Study 2 – the self-as-a-representational-base

So far, I showed that people could have different mental representations for different minimal groups, but to different extents for different category labels. If people do not have as clear a mental representation of Klee and Kandinsky compared to overestimators and underestimators, or if there is less consensus among different individuals about what Klee and Kandinsky fans look like, then how did people solve the face categorization task in Study 1? Did they simply choose faces that looked more favorable for ingroup and less favorable for outgroup, supporting the evaluative matching hypothesis? This would indicate that ingroup positivity is the only top-down influence on people's face representations when category labels are not particularly meaningful (i.e., low inductive potential). Could there be other sources of top-down influence on face representations?

One possibility is that people may use their own self-image to mentally represent novel ingroup faces but not outgroup faces. In a typical minimal group context, participants have only two pieces of information available to them: (1) category label of the target (e.g., overestimator or underestimator) and (2) whether the target shares the same group membership with them or not (ingroup or outgroup). When people perceive an initial sense of similarity to the target, they may assume that the target thinks, feels, and behaves the way they themselves think, feel, and behave (Ames, 2004a, 2004b). Sharing a group membership

with a target (i.e., ingroup) may signal similarity to the target compared to not sharing a group membership (i.e., outgroup), leading to more projection (i.e., using the self as a basis for understanding ingroup but not outgroup). This in fact has been demonstrated in the literature in domains such as information processing and evaluations (Clement & Krueger, 2002; Gramzow et al., 2001, 2005; Holtz & Miller, 1985; Krueger & Zeiger, 1993). In line with the past research, I hypothesized that if participants have to visualize what their ingroup looks like they might may at least in part use their self-image as a basis for their visualization. More concretely, the self-as-representational bias hypothesis would predict that people would generate ingroup faces that look more like themselves in the face categorization task compared to outgroup faces.

Study 2a

Study 2a tested the self-as-a-representational base hypothesis using the participant-level classification images of ingroup and outgroup as well as the photographs of the participants who generated these images. In this study, I recruited an independent sample of participants. On each trial of the main task, participants saw a photograph along with pairs of classification images, one of which was generated by the person in the photograph. I instructed participants to pick one of the two classification images that most resembled the person in the photograph. If the self-as-a-representational base hypothesis is true, I expected to find higher accuracy for ingroup images than outgroup images.

Method

Participants. I recruited 171 University of California Santa Barbara students (M_{age} = 18.89, SD = 1.44; 118 female, 53 male) to participate in a study about matching faces in exchange for course credit via the UCSB Psychological and Brain Sciences subject pool. The

racial and ethnic breakdown of participants was 64 White, 32 Latinx, 50 Asian, 1 Black, 13 multiracial, 10 other, and 1 unidentified. Up to four participants were run simultaneously.

Stimuli. I used 289 (out of 362) photographs of participants who generated ingroup and outgroup face images and provided a separate consent to be photographed from Study 1a (overestimator versus underestimator). The distribution of group memberships was 143 who generated ingroup images and 146 who generated outgroup. images Although I did not have an equal number of photographs of participants in each category, a chi-square goodness-of-fit test showed that there is no difference between ingroup and outgroup numbers, χ^2 (1, N = 289) = .03, p = .86.

Procedure. After providing consent, participants completed a task where they matched faces that looked most like each other. On each trial, participants saw three images: the top image was a photograph of a person who participated in the face categorization task in Part 1, and two bottom participant-level classification images, masked to only show face regions (i.e., background and hair removed), one of which was generated by the person in the photograph and the other was randomly chosen from the pool of 288 (289 - 1) participant level classification images generated by one of 288 remaining participants whose pictures were used in other trials (see Figure 9). Participants were then asked to indicate which of the bottom two images looked most like the photograph on top regardless of the gender of the photograph because all participant-level classification images were male faces (generated from a male base face image); they pressed the ‘d’ key if they thought the image on the left looked most like the photograph and the ‘k’ key if they thought the image on the right looked most like the photograph.

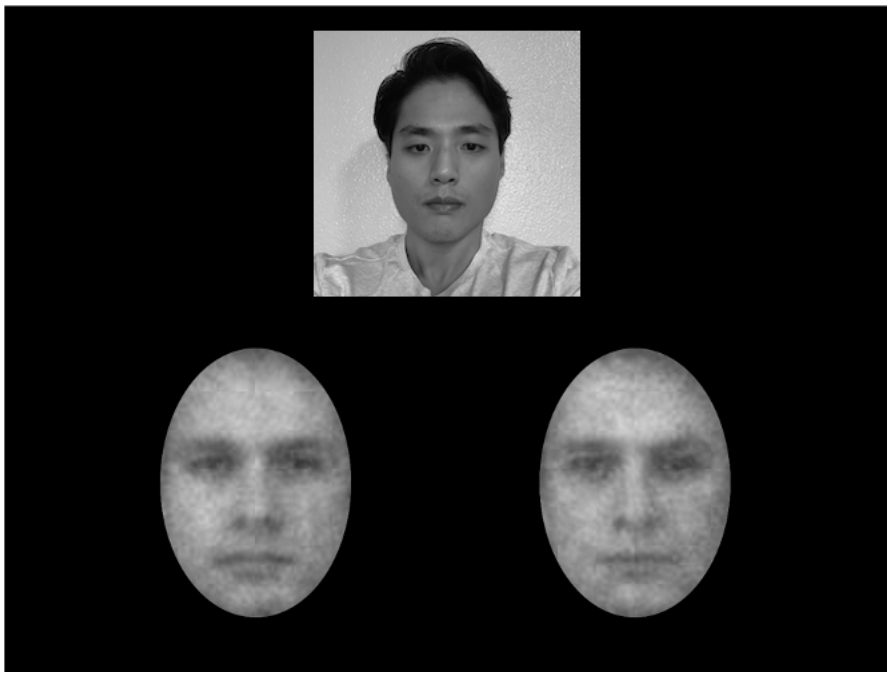


Figure 9. Photograph matching example. The photograph in this image is not of an actual participant to protect the anonymity of the participants.

Results

One-sample t-tests. First, I tested the hypothesis that people use their self-image to visually represent ingroup faces more than outgroup faces by conducting one-sample t-tests to see if participants in this study performed better than chance (50%) for ingroup and outgroup images. The participants correctly matched ingroup participant-level classification images with photographs of people who generated the images significantly better than chance (M = 51.77%, SD = 7.98%), t(142) = 2.65, p < .01, 95% CI [50.45, 53.09], Cohen's d = .22. On the other hand, matching of outgroup participant-level classification images with photographs of the people who generated them did not significantly differ from chance level (M = 49.32%, SD = 8.15%), t(145) = -1.02, p = .31, 95% CI [47.98, 50.65], Cohen's d = .08.

Mean-level accuracy comparison using linear mixed effects model. I also used the linear mixed effects model to compare the mean level difference in photo matching accuracy

performance between ingroup and outgroup. The model predicted a binomial outcome (0 = incorrect; 1= correct) using logistic regression and allowed the intercept and slope of group to vary. I found a main effect of group membership, χ^2 (1) = 29.37, $p < .001$, indicating that the accuracy was significantly higher for ingroup than for outgroup participant-level classification images. The estimated accuracy for ingroup was 51.8%, 95% CI [51.1, 52.4] and for outgroup was 49.3%, 95% CI [48.7, 49.9]. Because every classification image was based on an average white male face, I ran a separate model with race and gender of participants in the photograph as covariates to rule out the possibility that only the white male participants' images showed significant results. The main effect of group membership remained significant, χ^2 (1) = 25.15, $p < .001$, the estimated accuracies: ingroup = 51.5%, 95% CI [50.5, 52.6] and outgroup = 49.3%, 95% CI [48.2, 50.3].

Discussion

Study 2a examined the possibility that people use their self-image to visually represent minimal ingroup faces. The results of one sample t-tests showed that the participants could match ingroup classification images with photographs of people who generated them statistically better than chance, whereas the same was not true for outgroup images, supporting the self-as-a-representational-base hypothesis. I also found that overall ingroup images resulted in higher accuracy than outgroup images even after controlling for participant race and gender, providing evidence that people used self-image more for visualizing ingroup faces than outgroup faces, regardless of their own gender or race.

One possible alternative explanation of the current findings is that because ingroup faces were more trustworthy looking than outgroup faces (see Analysis 1 of Study 1a) and because trustworthiness is a unique human trait (Wilson et al., 2018), participants in this

study simply matched photographs to more human-looking (i.e., more trustworthy-looking) classification images, leading to greater accuracy scores for ingroup faces. Therefore, examining the trustworthiness of participant-level classification images to rule out this alternative hypothesis might be necessary (see Study 3). It is important to note that although we did find some evidence of people using their self-image as a representational base given the significantly better than chance results for ingroup, the effects were relatively small (1~2% better than chance), indicating that even though the self could have had a top-down influence on mental representations of minimal ingroup and outgroup faces, other factors such as ingroup positivity and category labels might have had a greater influence.

Study 2b

Study 2a provided initial evidence of self-as-a-representational base by showing that people were more likely to use their self-image to mentally represent minimal ingroup faces than outgroup faces. Study 2b attempted to replicate these findings with a different type of minimal group paradigm.

Method

Participants. I recruited 148 University of California Santa Barbara students (M_{age} = 19.16, SD = 1.44; 94 female, 54 male) to participate in a study about matching faces in exchange for course credit via the UCSB Psychological and Brain Sciences subject pool. The racial and ethnic breakdown of participants was 50 Asian, 38 Latinx, 38 White, 16 multiracial, 3 Black, 1 pacific islander/Hawaiian, 2 other, and 1 unidentified. Initially, I had planned to recruit a comparable number of participants to Study 2a (n = 171), but due to the COVID-19 pandemic and the subsequent in-person experiment closure at UCSB, I ended up with the current sample size. Up to four participants were run simultaneously.

Stimuli. I used the 129 (out of 200) photographs of participants who generated ingroup and outgroup face images and provided their consent to be photographed from Study 1b (Klee versus Kandinsky). The breakdown of the group memberships was 67 ingroup and 62 outgroup. Although I did not have an equal number of photographs of participants in each category, a chi-square goodness-of-fit test showed that there is no difference between ingroup and outgroup numbers, χ^2 (1, N = 129) = .19, p = .66.

Procedure. As was the case for Study 2a, participants completed a task where they matched faces that looked most like each other (see the Method section of Study 2a). The only difference was that Study 2b had 129 trials as opposed to 289 trials in Study 2a.

Results

One-sample t-tests. First, I tested the main hypothesis that people would use their self-image to represent their ingroup faces more than outgroup faces by conducting one-sample t-tests to see if participants performed better than chance (50%) for ingroup and outgroup images. The participants' matching accuracy for ingroup participant-level classification images and photographs of people who generated the images did not significantly differ from chance (M = 51.42%, SD = 8.69%), t(66) = 1.34, p = .19, 95% CI [49.30, 53.54], Cohen's d = .16. Additionally, their matching accuracy of outgroup participant-level classification images with photographs of the people who generated them was significantly worse than chance (M = 46.84%, SD = 8.70%), t(61) = -2.86, p < .01, 95% CI [44.63, 49.05], Cohen's d = .36.

Mean-level accuracy comparison using multilevel modeling. Next, I used the linear mixed effects model to compare the mean level difference in photo matching accuracy performance between ingroup and outgroup. The model predicted a binomial outcome (0 =

incorrect; 1= correct) using logistic regression and allowed the intercept and slope of group to vary. I found a main effect of group membership, χ^2 (1) = 39.43, p < .001, AIC = 26075.4, indicating that participants matched ingroup images to the photographs of people who generated them significantly better than for outgroup images. The estimated accuracy for ingroup was 51.4%, 95% CI [50.4, 52.4] and for outgroup was 46.8%, 95% CI [45.8, 47.9]. Because every classification image was based on an average white male face, I ran a separate model entering race and gender of participants in the photograph as covariates to rule out the possibility that only the white male participants' images showed significant results. The main effect of group membership remained significant, χ^2 (1) = 43.53, p < .001, although the estimated accuracy for ingroup was no longer better than chance: ingroup = 50.9%, 95% CI [49.4, 52.3] and outgroup = 45.8%, 95% CI [44.2, 47.3].

Discussion

Study 2b failed to fully replicate the findings from Study 2a using a different minimal group paradigm. Using one-sample t-tests, I did not find evidence of people using self-image to mentally represent their ingroup faces. However, I showed that people might have actively avoided using their self-image to mentally represent outgroup faces. The liner mixed-effects model, however, replicated the findings of Study 2a as indicated by significantly greater matching accuracy for ingroup images than for outgroup face images. I also found that overall ingroup images resulted in higher accuracy than outgroup images even after controlling for participant race and gender, although this effect was driven by significantly worse than chance accuracy for outgroup rather than significantly better than chance accuracy for ingroup.

The same limitations of Study 2a apply to the current study. The alternative hypothesis I suggested in Study 2a that participants simply chose faces that looked more trustworthy as resembling the people in the photographs could apply here. This could explain the significantly worse than chance accuracy for outgroup images given a big difference in trustworthiness found between ingroup and outgroup faces in the Klee/Kandinsky paradigm. That is, because the outgroup images were significantly less trustworthy looking than ingroup images in this study, participants particularly struggled to match them to the photographs of people who generated them. Therefore, in Study 3 I discuss two studies that assessed the perceived trustworthiness of participant-level classification images to further examine this alternative explanation.

Study 3 – the self-as-an-evaluative-base

Although in Study 2, I provided initial evidence that people may use their self-image when visualizing novel ingroup faces, the possibility remains that the findings were simply driven by people choosing more "human-looking" faces as resembling the people in the photographs. It is especially likely given that ingroup images were rated higher on uniquely human traits (e.g., trustworthiness; Wilson et al., 2018) than outgroup images in Study 1a and 1b. However, both results were based on ratings of group-level classification images. Therefore, in Study 3 I conducted two studies examining how much trusting behavior ingroup and outgroup participant-level classification images from Study 1 elicit in others. To this end, I recruited independent samples of participants to play an economic trust game with various partners represented by the ingroup and outgroup participant-level classification images from Study 1a and 1b (Berg, Dickhaut, & McCabe, 1995; Study 3 of Ratner et al., 2014). By doing so, I could test whether perceived trustworthiness of each image moderates

how accurately they were matched to the photographs of people who generated them. If the alternative explanation of the findings of Study 2 is true, then more trustworthy looking images would be more accurately matched to the photographs of people who generated them than less trustworthy looking images, regardless of whether they are ingroup or outgroup.

Additionally, collecting trustworthiness of every participant-level classification image allowed me to show another converging evidence of top-down influence of ingroup positivity on face representations. As the group-level ingroup images were rated as more trustworthy than the outgroup counterparts in Study 1, I expected to find similar results: ingroup participant-level images would be trusted more in the economic trust game than outgroup participant-level images. Such results would be consistent with past minimal group research showing the ingroup favoritism bias in evaluations and monetary allocations (e.g., Brewer & Silver, 1978; Tajfel et al., 1971). However, another interesting question can be explored with such a dataset. Is everyone equally motivated to represent ingroup members more favorably than outgroup members or do characteristics of the perceiver also matter? It is clear from previous research that individual differences play a role when it comes to judging trustworthiness from faces (Mattarozzi et al., 2015). Furthermore, moderating effects of individual differences on ingroup positivity outside of face representations have been documented, such as personal self-esteem (Rosenberg, 1965), collective self-esteem (Luhtanen & Crocker, 1992), and many more. Thus, it is reasonable to expect similar moderating effects of these variables on ingroup positivity in face representations. Measuring trustworthiness of each participant-level image allows for examining continuous relationships between the extent to which each image was trusted and individual differences of people who generated those images, and how these relationships differ between ingroup

and outgroup. Specifically, I examined the effects of following variables on people's face representations of minimal ingroup and outgroup.

Personal self-esteem. Previous research showed that people extend their positive views about themselves to their ingroups but not outgroups (Crocker et al., 1987; Cadinu & Rothbart, 1996; Clement & Krueger, 2002). For example, Gramzow and Gaertner (2005) showed that personal self-esteem (PSE) as measured by Rosenberg Self-Esteem scale (1965) was related to increased positivity toward novel ingroup and decreased positivity toward novel outgroup. I similarly predicted that PSE would be related to more trustworthy ingroup face representations as indicated by more money received in the trust game. On the other hand, PSE would be related to less trustworthy outgroup face representations as indicated by less money received in the trust game.

Collective self-esteem. In a similar vein, previous research showed that (private) collective self-esteem (CSE), defined as "the extent to which individuals *generally* evaluate their social groups positively" (Luhtanen & Crocker, 1992), is related to ingroup favoritism bias in general evaluation (Crocker & Luhtanen, 1990; Gramzow & Gaertner, 2005). That is, people high in CSE rated ingroup members more positively than outgroup members, but people low in CSE did not. Again, I expected to find a similar pattern of results: CSE would be related to more trustworthy ingroup face representations, whereas CSE would not be related to outgroup facial trustworthiness representations.

In invoking the moderating effects of personal self-esteem in face representation of ingroup and outgroup, I am are squarely in line with research that has shown that people use their self-evaluation as a basis for evaluating their ingroup members (Gramzow & Gaertner, 2005; Park & Judd, 1990). That is, in contrast to the self-as-a-representational-base

hypothesis tested in Study 2, Study 3 tested the self-as-an-evaluative-base hypothesis by examining the role of the self-evaluation on how people visually represent faces of minimal ingroup and outgroup members.

Study 3a

Methods

Participants. I recruited one hundred and eight 108 University of California, Santa Barbara students (M_{age} = 18.77, SD = 1.37; 67 female, 41 male) to participate in an economic trust game study in exchange for course credit via UCSB Psychological and Brain Sciences subject pool. Racial and ethnic breakdown of our sample was 40 White, 32 Asian, 20 Latinx, 1 Black, 10 multiracial, and 5 other. Up to four participants were run simultaneously.

Procedure. After consent, participants played an economic trust game with different interaction partners. The interaction partners were the ingroup and outgroup face classification images from Part 1. I instructed participants to imagine that they had $10 on each trial, and that they could choose either to keep this money or to share a certain amount of money with their interaction partners. On each interaction trial, participants made a choice to share a portion of $10 (i.e., $0, $2, $4, $6, $8, or $10). Participants were told that any money they shared would be quadrupled and given to the interaction partner. The interaction partner would then have the option to return half of the sum to the participant who had shared the money. In this way, it would be possible for the participant to make more money than if they had not shared. Participants simply indicated how much money they would like to share with each partner and did not receive any feedback (see Figure 10). The amount of money shared was indicative of extent to which the participants trusted the interaction partners. Participants played the trust game with a total of 362 different partners, represented only by a

classification image Study 1a. Different classification images were presented in a random order. The study concluded with a debriefing.

Figure 10. Trust game example. Participants moved mouse cursor to indicate how much money they would like to share with each interaction partner.

Results

I used the linear mixed effects model to test whether individual differences of the people who generated the ingroup and outgroup classification images were related to perceived trustworthiness of the classification images as indicated by how much money they received. Each interaction partner image had group membership (ingroup, outgroup) and individual differences scores (PSE and CSE) associated with it. The two individual differences scores were analyzed in two separate models. In each model, I entered group membership (ingroup, outgroup), individual differences scores, and their interaction term predicting how much money each image received. In all models the intercept and slope of group to vary, and treated participants in this study as a random factor.

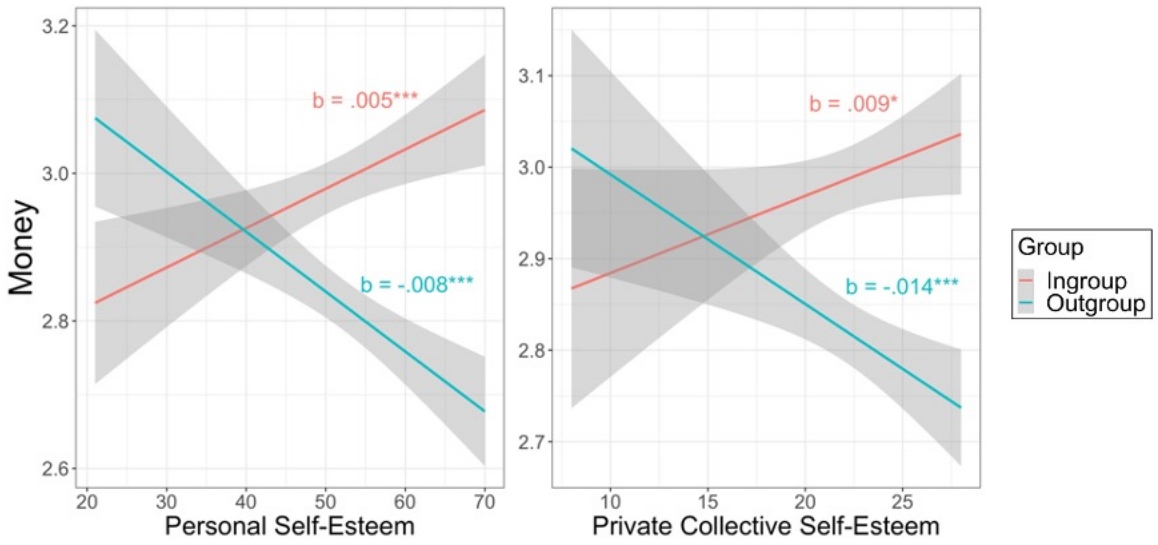
Ingroup favoritism. Before examining the effects of individual differences, I first examined whether there is overall ingroup favoritism bias (i.e., did ingroup images receive more money than outgroup images?). I found a main effect of group membership, χ^2(1) = 61.57, b = .16, p < .001, 95% CI [.121, .201]. That is, overall ingroup classification images received more money (M = 2.98, SD = 2.46) than outgroup classification images (M = 2.82, SD = 2.45).

Personal self-esteem. I mean-centered the personal self-esteem (PSE) score (M = 51.53 out of 70, SD = 9.62) for an easier interpretation. I found a main effect of group membership, $\chi^2(1) = 60.07$, b = .16, p < .001, 95% CI [.119, .199]. The main effect of PSE for ingroup was also significant, $\chi^2(1) = 15.08$, b = .005, p < .001, 95% CI [.003, .008], indicating that PSE was associated with more trustworthy ingroup faces. The interaction term was also significant, $\chi^2(1) = 46.31$, b = .013, p < .001, 95% CI [.010, .017], indicating that the slopes for ingroup and outgroup were significantly different from each other (see Figure 11). The simple slope for outgroup was significant, $\chi^2(1) = 31.77$, b = -.008, p < .001, 95% CI [-.005, -.011], indicating that PSE was associated with less trustworthy outgroup faces.

Private collective self-esteem. Because I was interested in how each individual generally evaluates their own groups (not specific to the minimal groups to which they were assigned), I focused on private CSE out of four subscales of CSE. I mean-centered the private CSE score (M = 21.91 out of 28, SD = 3.81). I found a main effect of group membership, $\chi^2(1) = 61.20$, b = .16, p < .001, 95% CI [.120, .201]. The main effect of private CSE for ingroup was significant, $\chi^2(1) = 5.68$, b = .008, p = .02, 95% CI [.001, .015], indicating that private CSE was associated with more trustworthy ingroup faces. The interaction term was also significant, $\chi^2(1) = 20.64$, b = .023, p < .001, 95% CI [.013, .032],

indicating that the slopes for ingroup and outgroup were significantly different from each other (see Figure 11). The simple slope for outgroup was significant, $\chi^2(1) = 16.43$, $b = -.014$, $p < .001$, 95% CI [-.007, -.013], indicating that private CSE was associated with less trustworthy outgroup faces.

Next, given that private CSE and PSE were significantly correlated with each other, $r(360) = .45$, $p < .001$, I conducted a separate analysis to examine whether PSE and CSE are independent predictors of trustworthiness of ingroup and outgroup participant-level classification images. To do this, I entered group membership, PSE, private CSE, and interaction terms between group membership and each self-esteem score predicting how much money each image received. I found that PSE and the group X PSE interaction term remained significant, $p < .01$, whereas the effects of private CSE and the group X CSE interaction term became no longer significant, $p > .05$, indicating that private CSE did not predict trustworthiness of ingroup and outgroup face images above and beyond PSE



Significance codes: *** <.001 ** <.01 * <.05 $^{+}$<.10

Figure 11. Study 3a moderating effects of PSE and private CSE on ingroup and outgroup facial trustworthiness representations.

Reanalysis of Study 2a data. Another main purpose of the current study was to examine the possibility that in Study 2a, participants simply matched faces that looked more trustworthy looking to the photographs of people who generated them. The trust game data from the current study can partially address that issue. I did so by first computing the difference score between money received for the correct image (i.e., the image that the person in the photograph generated) and money received for the distractor image (i.e., the image that the person in the photograph did not generate) for every trial from Study 2a. If people indeed chose more trustworthy looking faces, then as the difference in money received between the correct image and the distractor image increases in favor of the correct image (i.e., the correct image is more trustworthy looking), they should have been matched with higher accuracy. To test this hypothesis, I entered group membership (ingroup, outgroup) as a predictor and trust game money difference score as a covariate predicting

photo matching accuracy in a linear mixed effects model. Same as before, the model predicted a binomial outcome (0 = incorrect; 1= correct) using logistic regression and allowed the intercept and slope of group to vary. The main effect of group membership remained significant, χ^2 (1) = 9.29, p = .002, indicating that the accuracy was still significantly higher for ingroup than for outgroup participant-level classification images. The estimated accuracy for ingroup was 51.2%, 95% CI [50.6, 51.9] and for outgroup was 49.8%, 95% CI [49.2, 50.5]. The main effect of trust game money difference score was also significant, χ^2 (1) = 117.04, b = .188 p < .001, 95% CI [.169, .208], meaning that as the correct image's perceived trustworthiness (money received in the trust game) increased relative to the distractor image's perceived trustworthiness, it was matched to the photograph of the person who generated the image with higher accuracy (see Figure 12).

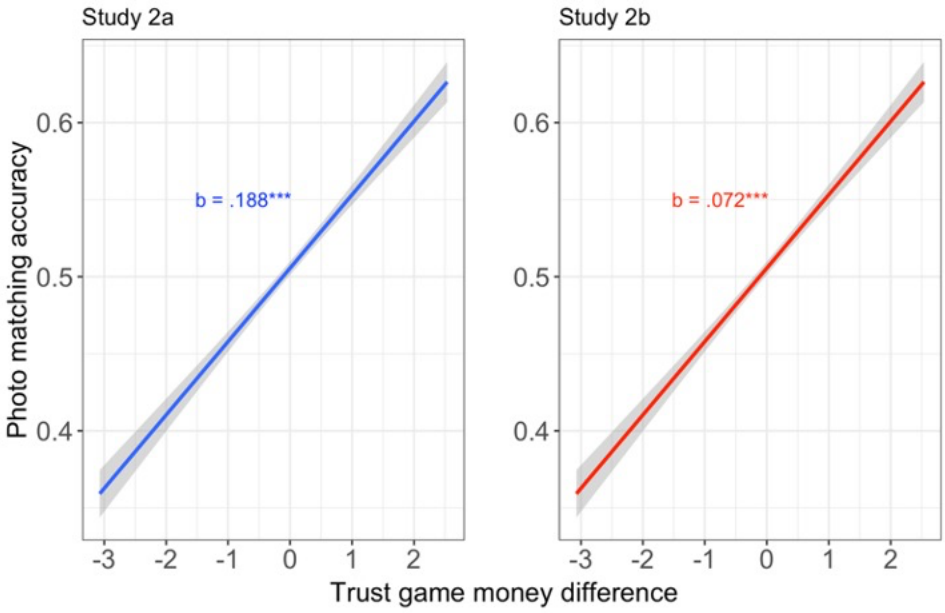


Figure 12. The relationship between perceived trustworthiness difference (between correct image and distractor image) and photo matching accuracy.

Discussion

The purpose of Study 3a was two-fold. First, I examined the effects of two individual difference variables on ingroup positivity bias. I showed that personal self-esteem uniquely moderated perceived trustworthiness of ingroup and outgroup participant-level classification images from Study 1a. Overall, people showed an ingroup positivity bias: people generated ingroup face images that are more trustworthy than outgroup face images as indicated by more money received by ingroup faces than outgroup faces in the economic trust game. Furthermore, people high in personal self-esteem generated even more trustworthy ingroup faces and even less trustworthy outgroup faces. Private collective self-esteem had similar effects, but it was not uniquely related to perceived trustworthiness of ingroup and outgroup face images after controlling for personal self-esteem. The finding that only personal self-esteem uniquely predicted perceived trustworthiness of their face representations of ingroup and outgroup is consistent with the self-as-an-evaluative base findings from previous studies demonstrating that personal self-esteem is related to evaluations of minimal ingroup and outgroup above and beyond collective self-esteem (Gramzow & Gaertner, 2005).

Second, I examined the alternative explanation for the self-as-a-representational base findings from Study 2a by controlling for perceived trustworthiness (i.e., money received in the trust game) of each participant-level classification image obtained from the current study in the reanalysis of Study 2a data. I found that even after controlling for perceived trustworthiness of classification images, ingroup face images were matched to the photographs of people who generated them with higher accuracy than outgroup face images, and in fact the ingroup images were still matched better than chance. However, I also found a significant effect of perceived trustworthiness of participant-level classification images: the

more trustworthy corrected images looked relative to distractor images, the more accurately they were matched to the photographs of people who generated them. This finding partially supports the alternative explanation that people simply chose more trustworthy looking images in the photo matching task in Study 2a, and since overall ingroup faces looked more trustworthy, they were matched with higher accuracy. Although I could not rule out this possibility, it is important to note that the main effect of group membership remained significant in this analysis. It is possible that the self-as-a-representational base perspective still holds true and that the inability to rule out alternative explanations (i.e., confounding effects of perceived trustworthiness) is simply attributed to the design flaws of the photo matching task from Study 2. Thus, future research should develop alternative methods for comparing photographs to classification images and rule out any extraneous influences such as perceived trustworthiness of classification images.

Study 3b

Study 3a demonstrated the effects of personal and collective self-esteem on ingroup positivity in face representation of ingroup and outgroup by showing that people with higher self-esteem generated ingroup faces that were trusted more and outgroup faces that were trusted less. Study 3a also provided some support for the alternative explanation of Study 2a findings that ingroup face images were matched with higher accuracy simply because they were more trustworthy looking than outgroup face images. Study 3b attempted to replicate these findings with participant-level classification images from Study 2b.

Method

Participants. I recruited 148 University of California, Santa Barbara students (M_{age} = 19.16, SD = 1.44; 94 female, 54 male) to participate in a study about face games in exchange

for course credit via the UCSB Psychological and Brain Sciences subject pool. They are the same participants who completed the photo matching task in study 2a. The order for the two tasks (photo matching task and trust game) were randomly determined across participants. Racial and ethnic breakdown of participants was 50 Asian, 38 Latinx, 38 White, 16 multiracial, 3 Black, 1 Pacific Islander/Hawaiian, 2 other, and 1 unidentified. Up to four participants were run simultaneously.

Procedure. As was the case for of Study 3a, participants played an economic trust game. with different interaction partners. The interaction participants in this study were represented by the ingroup and outgroup participant-level classification images from Study 1b. The only difference was that Study 3b had 200 trials instead of 362 trials in Study 3a.

Results

Again, I used the linear mixed effects model to test whether the self-esteem scores of the people who generated the ingroup and outgroup classification images were related to perceived trustworthiness of the classification images as indicated by how much money they received. Each image had group membership (ingroup, outgroup) and individual differences scores (PSE and CSE) associated with it. The two individual differences scores were analyzed in two separate models. In each model, I entered group membership (ingroup, outgroup), individual differences scores, and their interaction term predicting how much money each image received. In all models the intercept and slope of group to vary, and treated participants in this study as a random factor.

Ingroup favoritism. Before examining the effects of individual differences, I first examined whether there is overall ingroup positivity bias. I found a main effect of group membership, $\chi^2(1) = 361.62$, $b = .93$, $p < .001$, 95% CI [.83, 1.02]. That is, overall ingroup

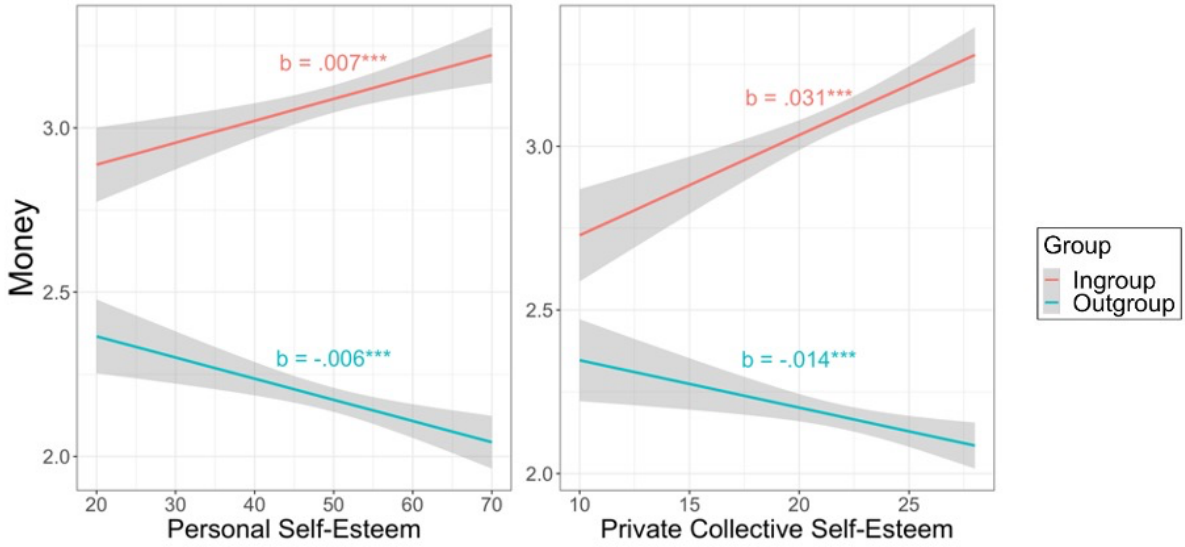
classification images received more money (M = 3.10, SD = 2.55) than outgroup classification images (M = 2.17, SD = 2.25).

Personal self-esteem. I mean-centered the personal self-esteem (PSE) score (M = 49.63 out of 70, SD = 11.12) for an easier interpretation. I found a main effect of group membership, $\chi^2(1) = 357.22$, b = .91, p < .001, 95% CI [.82, 1.01]. The main effect of PSE for ingroup was also significant, $\chi^2(1) = 23.59$, b = .007, p < .001, 95% CI [.004, .009], indicating that PSE was associated with more trustworthy ingroup faces. The interaction term was also significant, $\chi^2(1) = 40.52$, b = .013, p < .001, 95% CI [.009, .017], indicating that the slopes for ingroup and outgroup were significantly different from each other (see Figure 13). The simple slope for outgroup was significant, $\chi^2(1) = 17.60$, b = -.006, p < .001, 95% CI [-.003, -.009], indicating that PSE was associated with less trustworthy outgroup faces.

Private collective self-esteem. I mean-centered the private CSE score (M = 21.86 out of 28, SD = 3.68). I found a main effect of group membership, $\chi^2(1) = 359.22$, b = .91, p < .001, 95% CI [.82, 1.01]. The main effect of private CSE for ingroup was significant, $\chi^2(1) = 47.71$, b = .03, p <.001, 95% CI [.022, .039], indicating that private CSE was associated with more trustworthy ingroup faces. The interaction term was also significant, $\chi^2(1) = 53.32$, b = .05, p < .001, 95% CI [.033, .057], indicating that the slopes for ingroup and outgroup were significantly different from each other (see Figure 13). The simple slope for outgroup was significant, $\chi^2(1) = 11.34$, b = -.014, p < .001, 95% CI [-.006, -.023], indicating that private CSE was associated with less trustworthy outgroup faces.

Next, given that private CSE and PSE were significantly correlated with each other, $r(196) = .42$, p < .001, I conducted a separate analysis to examine whether PSE and CSE are independent predictors of trustworthiness of ingroup and outgroup participant-level

classification images. I entered group membership, PSE, private CSE, and interaction terms between group membership and each self-esteem score predicting how much money each image received. I found that PSE, the group X PSE interaction term, private CSE, and the group X CSE interaction all remained significant, $p < .01$, indicating that PSE and private CSE each uniquely predicted trustworthiness of ingroup and outgroup face images.



Significance codes: *** <.001 ** <.01 * <.05 + <.10

Figure 13. Study 3b moderating effects of PSE and private CSE on ingroup and outgroup facial trustworthiness representations.

Reanalysis of Study 2b data. To address the alternative explanation of the Study 2b finding that people matched more trustworthy looking faces to the photographs, hence the higher accuracy for ingroup faces than outgroup faces, in a separate analysis I entered group membership (ingroup, outgroup) as a predictor and trust game money difference score as a covariate predicting photo matching accuracy in a linear mixed effects model. The model predicted a binomial outcome (0 = incorrect; 1= correct) using logistic regression and allowed the intercept and slope of group to vary. The main effect of group membership

remained significant, χ^2 (1) = 12.19, p < .001, indicating that the accuracy was still significantly higher for ingroup than for outgroup participant-level classification images. The estimated accuracy for ingroup was 50.6%, 95% CI [49.5, 51.6] and for outgroup was 47.8%, 95% CI [46.7, 48.8]. The main effect of trust game money difference score was also significant, χ^2 (1) = 25.77, b = .07 p < .001, 95% CI [.04, .10], meaning that as the correct image's perceived trustworthiness (money received in the trust game) increased relative to the distractor image's perceived trustworthiness, it was matched to the photograph of the person who generated the image with higher accuracy (see Figure 13).

Discussion

Study 3b attempted to replicate the findings of Study 3a using participants-level classification images from a different minimal group paradigm. First, I replicated the ingroup positivity finding: overall, people generated ingroup faces that were trusted more than outgroup faces. Second, I showed that both personal and collective self-esteem uniquely moderated top-down influences of ingroup positivity bias on people's face representations of minimal ingroup and outgroup. People high in self-esteem generated even more trustworthy ingroup faces and even less trustworthy outgroup faces than people low in self-esteem. Third, I examined the alternative explanation for the self-as-a-representational-base findings from Study 2b by controlling for perceived trustworthiness (i.e., money received in the trust game) of each participant-level classification image obtained from the current study in the reanalysis of Study 2b data. I found that even after controlling for perceived trustworthiness of classification images, ingroup face images were matched to the photographs of people who generated them with higher accuracy than outgroup face images. This effect was driven by significantly worse than chance accuracy for outgroup face images unlike significantly

better than chance accuracy for ingroup face images in Study 2a. I also found a significant effect of perceived trustworthiness of participant-level classification images: the more trustworthy correct images looked relative to distractor images, the more accurately they were matched to the photographs of people who generated them. Again, this finding provides another evidence for the possibility that people simply chose more trustworthy looking images in the photo matching task. Although it is difficult to completely dismiss the interpretation that people used their self-image to visually represent minimal ingroup faces but not outgroup faces, it appears likely that the use of self-image was not a prominent driver of how people visually represented faces of ingroup members. Nonetheless, in this re-analysis, the main effect of group membership remained significant (after controlling for perceived trustworthiness of the images), suggesting that the alternative explanation in and of itself could not explain the results.

General discussion

Despite the popularity of the minimal group paradigm and its long history, no previous studies have rigorously tested the foundational assumption that minimal group labels have no consequential meaning to participants. The goal of this dissertation was to empirically examine whether the classic minimal group paradigms truly remove top-down influences of knowledge structure on face representations. Specifically, I examined top-down influences of minimal group category labels and self-knowledge on face representations of ingroup and outgroup in this research.

I first used the numerical estimation style minimal group paradigm to replicate Ratner et al.'s (2014) finding that people form mental representations of ingroup faces that are associated with more favorable traits than are outgroup face representations in Study 1a.

Beyond successfully replicating their ingroup positivity main effect, I showed that category labels also mattered, in that, faces of overestimators and underestimators are represented differently. The initial evidence for these conclusions was drawn from statistically comparing whether the experimental conditions influenced the trait rating means of the classification images, which was a strategy used by Ratner et al. (2014) to examine differences between ingroup and outgroup as well as many other research using the reverse correlation method (Dotsch &. Todorov, 2012). I then conducted representational similarity analysis on the trait rating data, and found that the overestimator label contributed to the overall pattern of trait ratings more so than the ingroup/outgroup distinction, but this was not the case for the underestimator label. I further corroborated these conclusions with a machine learning analysis. The machine learning analysis suggested that the NEST labels (overestimator vs. underestimator) might in fact contribute more to the face representations than the ingroup/outgroup distinction as indicated by the algorithm's better accuracy for classifying the NEST labels than the ingroup/outgroup labels. Next, I sought to understand the generalizability of the findings from Study 1a by using a different minimal group paradigm (i.e., Klee vs. Kandinsky preference), which on its face has less inductive potential than does the overestimator/underestimator minimal group paradigm. I found that Klee and Kandinsky fans are represented differently at the group-level, consistent with the overestimator/underestimator results. However, the representational similarity analysis revealed that the Klee and Kandinsky labels contributed to the face representations equal to or less than the ingroup/outgroup distinction. This in turn was associated with larger differences between ingroup and outgroup in both trait ratings and machine learning analyses.

I also examined another possible source of top-down influences on face representations of ingroup and outgroup – the self. In Study 2, I tested the self-as-a-representational-base hypothesis by examining whether people used their self-image to visually represent ingroup faces but not outgroup faces. I found partial evidence supporting this hypothesis: the ingroup images from the overestimator/underestimator paradigm were matched to the photographs of people who generated them significantly better than chance, whereas the outgroup images from the Klee/Kandinsky paradigm were matched to the photographs of people who generated them significantly worse than chance. These effects were present even after controlling for race and gender of the people who generated the images, which were based on an average white male face, indicating that people might use their self-image to visually represent novel ingroup faces (or actively avoid using self-image to represent outgroup faces). Although these findings are interesting, an alternative explanation likely: people in the photo matching tasks simply chose classification images that looked more "human" as resembling people in the photographs by selecting more trustworthy looking faces.

Study 3 examined this possibility and also test the self-as-an-evaluative-base hypothesis by examining the relationship between self-evaluation as measured by self-esteem and the extent to which they generated ingroup and outgroup faces that were trusted in an economic trust game. In two experiments, I found that self-esteem was positively related to how trustworthy ingroup faces looked and negatively related to how trustworthy outgroup faces looked. This effect was stronger for personal than collective self-esteem, indicating that people used their self-evaluation as a basis for showing ingroup positivity bias above and beyond their general evaluation of their groups. Collecting trust game data also allowed for

testing the alternative explanation for the study 2 findings. I found that when the correct images were more trustworthy looking than the distractor images in the photo matching task, they were matched with higher accuracy, rather than because the ingroup face images actually resembled the people in the photographs more than the outgroup face images.

What does this all mean for Tajfel et al.'s (1971) foundational assumption that the minimal group paradigm removes any top-down influence of knowledge structure? It appears that category labels are flimsy in some ways, but not in others. On one hand, ingroup positivity bias reliably emerges irrespective of category label. On the other hand, category labels are represented differently and can moderate ingroup positivity bias in representations. Moreover, when high in inductive potential, these labels can carry more weight in the overall representations than the ingroup/outgroup distinction, leading to less distinction between ingroup and outgroup and thus less ingroup positivity bias. Additionally, the self might be another top-down influence on the construction of mental representations of ingroup and outgroup faces, independent of ingroup positivity and category labels. Because of the limited information provided in a typical minimal group context, people may use their self-image to visualize members of their ingroup whom they have never met before. Alternatively, some people might be more motivated to express their ingroup positivity bias depending on how they evaluate themselves.

Lessons learned for minimal group research

Not all minimal groups are the same. This seems obvious, but this detail has been largely ignored in the minimal group literature. As Pinter and Greenwald (2010) point out, there has been oddly little development in group assignment techniques over the years. The paradigm has been successful in creating ingroup/outgroup distinctions and intergroup bias

that emerges from it, including in the current set of experiments. Moreover, the minimal group research has not suffered from replicability problems that have plagued other psychological paradigms. From a methodological standpoint, however, the current research suggests that researchers need to recognize that careful attention to counterbalancing and full reporting of category label differences is important to prevent interpretative slippage. Slight differences between category labels could lead to assuming ingroup versus outgroup differences that are really driven by the category labels. Because category label differences might not be the same from one version of the minimal group paradigm to another, it should not be assumed that they are all interchangeable and will cause exactly the same effects. Similarly, in an exploratory vacuum like the minimal group paradigm, people might have nowhere to rely on but the self to base their judgments, especially if they do not or cannot imbue meaning to category labels. I found evidence for this idea based on people using their self-image and self-evaluation to base their ingroup representations. Therefore, individual differences of participants in minimal group experiments must also be taken into account, especially when investigating face representations, which may be sensitive to various top-down influences beyond simple ingroup positivity bias.

My research also provides insight into how the implied meaningfulness of category labels influences face representations of ingroup and outgroup. A priori, it was not obvious whether reading into the labels should increase intergroup bias or attenuate it. On one hand, the field has gravitated toward using contrived group distinctions instead of purely random ones. This suggests that some inferential grist on the labels could be important for making participants view the novel categories as entitative groups. From this perspective, the overestimator/underestimator (NEST) paradigm should lead to a stronger ingroup positivity

bias than the Klee/Kandinsky (ART) paradigm because the inductive potential of the NEST labels provides a rationale to identify with one's ingroup. On the other hand, if one is reading into the NEST labels and imply characteristics (e.g., overestimator are confident and aggressive), then these associations could obscure the ingroup/outgroup distinction. The current results support this latter possibility. One implication is that if one is trying to manipulate minimal groups that are mostly devoid of meaning, then the Klee/Kandinsky paradigm might be a better option. However, many real-world novel groups actually have labels that imply characteristics. To the extent that a researcher is interested in modeling this dynamic in their experiments, then maybe the overestimator/underestimator paradigm is more appropriate. All of this said, the differences between these two paradigms should not be overstated. Despite their differences, they both generally support the claim that separating people into novel groups leads to ingroup positivity bias. Moreover, the effects of the self were similarly present in both paradigms.

It is also clear from this research that simply telling participants that groups do not differ cannot be relied upon. Overestimator and underestimator were represented differently more than ingroup and outgroup, even though I told participants that numerical estimation style was not related to any other cognitive tendencies or personality traits. Klee and Kandinsky were also represented differently at the group-level even though I gave participants the same instructions, albeit this distinction was represented to a lesser extent than the ingroup and outgroup distinction. Why would people ignore the instructions designed to constrain their inferences about the underlying essence of the groups? People are motivated to make meaning of the world around them and as Bruner taught us long ago, people *go beyond the information given*. It is also the case that priming could contribute to

the label effects. Affect and semantic misattribution research uses paradigms that instruct participants to not let a prime influence them, yet participants are still influenced by the prime (e.g., Imhoff et al., 2011a; Payne et al., 2005). In a related vein, perhaps telling someone not to read into the meaning of a category label backfires in the same way that people struggle not to think about a white bear. Ironic process theory of mental control (Wegner, 1994) argues that the act of telling people to ignore a concept makes them think about it more because it keeps the concept active in their minds.

It is still unclear as to why participants represent Kandinsky and Klee groups differently. There is no reason to assume meaningful differences between these groups, unlike is the case for overestimators and underestimators. Maybe stereotypes associated with the Kandinsky and Klee surnames and their respective ethnicities are automatically transferred to the groups even though it is not logical to assume that fans of the abstract painters share their ethnicities. It is also possible that sound symbolism plays a role. Sound symbolism assumes that vocal sounds and phonemes carry meaning (Köhler, 1929). Recent research suggests that more sonorant phonemes are associated with high emotionality, agreeableness, and conscientiousness, whereas names with voiceless stop phonemes are associated with high extraversion (Sidhu et al., 2019). Maybe the way that category labels are pronounced can bias trait inferences. Additional research is necessary to explore these possibilities. Alternatively, a simpler explanation for this finding is that it may just be a false positive because the significant differences between Klee and Kandinsky were only observed in the group-level classification images. Recent simulation studies show that trait ratings of group-level classification images may be prone to Type I error (Cone et al., 2020). Thus, I

remain cautious about interpreting the finding that faces Klee and Kandinsky groups were represented differently.

The current research only focused on the two most famous minimal group paradigms. It was beyond the scope of this research to catalog the effects of all possible minimal group labels. Although future studies are necessary to understand how broadly the current findings generalize, the initial evidence of category label effects suggests that a revision is necessary to account for how categorization occurs in minimal group settings. The minimal group paradigm emerged in the early 1970s when similarity-based models dominated cognitive psychology's understanding of categorization (e.g., Rosch & Mervis, 1975). Not surprisingly, categorization during the minimal group paradigm was largely thought to straightforwardly involve matching characteristics of the perceiver with characteristics of the target. From this vantage, the category label served as a vehicle for creating ingroup/outgroup distinctions and nothing more. However, cognitive psychology researchers outside of social psychology soon began to argue that similarity-based models were not adequate to explain categorization. They suggested that categorization may be "more like problem solving than attribute matching" (Medin, 1989). This so-called *theory- based* view of categorization suggested that perceivers take note of the attributes that correlate with category membership, but categorization ultimately results from generating an explanatory principle of how these attributes are interrelated (Murphy & Medin, 1985). This conceptual development was recognized by some social perception researchers who used the theory-based view of categorization to explain racial essentialism (Rothbart & Taylor, 1992). However, such insight was never integrated into theories designed to explain minimal group effects. Theory-based categorization could help minimal group researchers account for category label effects.

It is possible that the explanatory vacuum created by the lack of meaning attributed to the minimal group labels prompts participants to wonder what produces the categorical distinction and why exactly they (and other people) are in one category versus the other. Therefore, it should not be a surprise if their participants attempt to read into the category labels and infer meaning from them. My research showed another way participants could solve this conundrum, by extending the self to novel ingroup. Such framing may sound like meaning making of category labels and using self as a basis for making decision mutually exclusive. However, it is possible that both have separate or even interactive top-down influences on how people visually represent faces of minimal ingroup and outgroup and that people utilize both strategies to make sense of a minimal group situation. The current research did not examine under what circumstances people may opt for using one strategy over another and whether the process of using these strategies is deliberate or spontaneous (i.e., are people aware that they are imbuing meaning to category labels or using self-image to generate ingroup faces?). Evidence that people used the self-image and self-evaluation to guide their decisions in both paradigms suggests that the inductive potential of the labels neither facilitates nor hinders the effects of the self. However, future research should further explore these questions to deepen our understanding of the role of ingroup positivity, category labels, and the self in face representations of ingroup and outgroup.

The role of category labels in face presentations of minimal ingroup and outgroup

Many everyday social categories are not natural kinds, in the sense that they are human creations that started with a seemingly arbitrary distinction. Yet, as is the case with minimal groups created in the laboratory, group distinctions in the real world are often given labels that are not completely devoid of meaning. When people are assigned to groups in

organizations, those groups typically have team names that give them an identity. Sports teams, street gangs, nation states, and many other groupings have names that provide strong social significance, even though group membership is mostly a product of where you happen to have been born or currently live. Even military units that are officially defined in an arbitrary manner have nick-names. For instance, members of the 101st airborne division of the United States Army are called the "Screaming Eagles." Members of the 34th infantry division are called the "Red Bulls." These labels give their groups a specific character.

In contrary to conventional wisdom, group labels matter when making sense of novel groups, the arbitrary distinctions from which intergroup biases are thought to start might not always be arbitrary and could actually provide a grist to infer traits and power and status differences. People are active meaning-makers and associations that come to mind when processing category labels might be the impetus that gets the ball rolling down the hill toward entrenched biases. The possibility that the labels given to novel categories could have this effect has never been seriously considered in the literature. Yet doing so, makes clear that the labels that people use to categorize could have implications for understanding real-world groups.

The current research shows how visualizing group members might have real-world implications in face perception as well as general social perception. Predictive coding theories suggest that visual perception heavily relies on our expectations or visual predictions about the world (Friston, 2005). Relatedly, the predictive social perception model relies on the assumption that processing of social information is hypothesis driven, meaning that predictions about socially relevant information guide processing of new social information (Bach & Schenke, 2017; Den Ouden, Kok, & de Lange, 2012;). Perhaps face representations

of minimal ingroup and outgroup members serve as the "predictions" about what an ingroup or outgroup member looks like. That is, for example, people might expect a novel group member to look a certain way. In this situation, one might spontaneously visualize the other person and this visualization, even if not grounded in reality, could bias their evaluations and impressions. My work suggests that in such situations, people might grasp onto any knowledge that is tangible and then relate this knowledge to some concept that they have experienced. Once people connect a novel group distinction to a concept that they have experienced, then they can retrieve perceptual details from memory to populate their visualization. It is possible that established group distinctions that now have considerable social meaning, such as race, at least partially developed their meaning through such a process. For instance, skin color is one of the most salient indicators of racial group (Maddox, 2004). However, skin color is a physical attribute and does not have inherent social meaning. It is possible that associations that people made with light and dark contributed to how people initially formed mental representations about race. Such associations people make about arbitrary attributes and characteristics might lead to suboptimal information processing and can persist in the face of counter evidence (Sherman et al., 2005; Srull & Wyer, 1989). This last possibility is purely speculative, but highlights how forming theories of what causes group distinctions could affect how people mentally represent groups in everyday life.

Whether subtle cues like category labels have any influence in real-life scenarios largely depends on the broader context. Unlike in a laboratory setting, which is highly controlled, everyday life is complex. Many groups are embroiled in intractable conflicts, are marked by stigmatizing stereotypes, and are embedded in unequal power structures. In these

situations, category label effects likely contribute negligible variability to intergroup responses. However, in the absence of these weighty factors, seemingly trivial factors such as category labels could have a surprisingly disproportionate influence. Recent research by Levari et al. (2018) shows that as the prevalence of a stimulus decreases people seek out other ways to distinguish categories. For instance, they found that when threatening faces become infrequent, participants start to view neutral faces as threatening. In relation to our research, this work suggests that if the usual drivers of intergroup conflict are not a factor, people might latch onto any attributes that differentiate one group from the other. One interesting implication of this idea is that when groups are easily distinguishable by the labels, such as the overestimator and underestimator labels, they may not feel the need to distinguish them as us versus them and not engage in ingroup favoritism or outgroup derogation as much. On the other hand, when the labels do not provide much meaning (i.e., low inductive potential), people may feel the need to distinguish them by bolstering their ingroup positivity bias to highlight the ingroup/outgroup distinction. Most groups are not labeled by random numbers or letter strings. Their names contribute to their identity. Thus, category labels could deliberately or spontaneously provide associations that feed intergroup differentiation.

The role of the self in face representations of minimal ingroup and outgroup

We often join or discover new social groups. A freshman in college may join a sorority. An NBA player may get traded to a new team. An immigrant may become a naturalized U.S. citizen. How do people make sense of such novel groups? In these scenarios, individuals might already have extensive knowledge about the group as a whole, such their history and culture, as well as individual group members. This knowledge can guide their

evaluation and perception of their new group, including how they visually represent group members (e.g., Dotsch et al., 2008; Imhoff et al., 2011b; Mangini & Biederman, 2004). How can people visualize faces of novel group members about whom they know very little to none, such as in a minimal group situation? The only two pieces of clear information available to participants in this situation are category labels and whether the target shares the same group membership with them (ingroup) or not (outgroup). Thus, one obvious answer to this question is that people simply have a more idealized mental representation of ingroup faces than outgroup faces as documented by Ratner et al. (2014) and throughout the current research. Another possibility is that because sharing the same group membership with someone implies similarity, people might also assume similarity in other characteristics (Ames, 2004a, 2004b). Could this be extended to how people visually represent faces of novel ingroup members? My work provided initial evidence of people using their own self-image to visualize ingroup faces as well as actively avoid using self-image to visualize outgroup faces. However, the effects were small and inconsistent. Other factors such as ingroup positivity and category labels might have been a more prominent driver of constructing mental representations of minimal ingroup and outgroup faces. From this vantage, the self-as-a-representational-base finding might simply be viewed as a faint echo of an overlearned response: Because people have continually observed across multiple instances that they tend to look like their fellow group members, they have developed a strong self and ingroup appearance association that has some spontaneous top-down influence on mental representation of even minimal ingroup faces even though the minimal groups are not defined by physical characteristics. It is also possible that I simply underestimated the true effect size of people using self-image to visualize their fellow ingroup members because the

base image used to create classification images was an average White male face when most of the participants in both studies who generated the images were not. This aspect of the design could have reduced the sensitivity of the photo matching task to detect the effects for non-White male participants. It is also the case that matching actual photographs of participants and classification images generated by them is a very difficult task given the ambiguity of the classification images. Thus, future research should develop more sensitive methods for comparing photographs to classification images and probe the generalizability of the current findings.

Alternatively, another aspect of the self-knowledge people can use to guide their attitudes and perceptions of novel ingroups is how they evaluate themselves. Ingroup, by definition, includes the self. Thus, if people regard themselves highly, then they should also be motivated to visually represent their ingroup in an equally favorable light, because their ingroup includes them. This was in fact demonstrated with general evaluations of groups in previous research (Gramzow & Gaertner, 2005). In this research, I showed that people's personal self-esteem moderates perceived trustworthiness of their mental representation of ingroup and outgroup faces. That is, people with high self-esteem generated face representation of ingroup members that are even more trustworthy and outgroup members that are even less trustworthy than people with lower self-esteem. Also, showing individual differences in ingroup positivity bias in face representations of ingroup and outgroup provides a more nuanced look at the top-down influences of ingroup positivity: some individuals might be more motivated to have more biased mental representations of ingroup and outgroup members. The significant negative relationship between self-esteem and perceived trustworthiness of outgroup faces might suggest that people with high self-esteem

are also motivated to engage in outgroup derogation to elevate their ingroup status (Crocker et al., 1987), which in turn can be a source of positive self-view (Tajfel & Turner, 1979), creating a feedback loop of self-esteem and outgroup derogation. This possibility is speculative at this stage, and it is unclear as to why people with high self-esteem generated outgroup faces that were more negative in minimal group contexts. Thus, future research should further examine conditions under which self-esteem is related to ingroup positivity and outgroup negativity in face representations.

Whether it is assuming that other group members also look like them or extending evaluations about the self to evaluations about their ingroup, people may use the self as a basis for visualizing faces of minimal ingroup and outgroup members. Although this work provided a proof-of-principle that the self has a top-down influence on face representations of minimal ingroup and outgroup, there remain many outstanding questions. For example, are some people more likely to use their own self-image to guide their visualization of ingroup faces than others? Perhaps people who perceive greater overlap between the self and their ingroup (e.g., Aron, Aron, & Smollan, 1992) might be more likely to think that fellow group members also look like them. That being said, the self clearly plays an important role in how people visually represent faces of ingroup and outgroup members, and future studies further examining the effects of the self and individual difference could also enrich our understanding of the minimal group paradigm and intergroup dynamics that result from it.

Significance for reverse correlation research in social psychology

The goal of the current research was to understand various top-down influences on face representations of minimal ingroup and outgroup. Because the reverse correlation method is sensitive to category representation it seemed like an ideal tool to use in the current

research. In the process of maximizing the utility of this method to explore the role of ingroup positivity, category labels, and the self in face representations of minimal ingroup and outgroup, I developed several novel ways of analyzing reverse correlation data. First, I introduced using machine learning to analyze participant-level classification images. Assessing participant-level classification images with traditional social psychological methods can be challenging because of the number of trials that are often needed. Although large numbers of participant-level classification images can take a long time to process with machine learning methods, this processing time is easier to manage than the fatigue and boredom of processing all of these images for human participants. Thus, the machine learning technique I used could prove useful to other social psychologists addressing different research questions with reverse correlation methods. Moreover, examining participant-level classification images also circumvents the problem of inflated Type I error rate from examining group-level classification images (Cone et al., 2020). Another aspect the machine learning analysis, feature selection to examine common structures underlying ingroup/outgroup distinction and trustworthy judgments was another novel way to answer questions that conventional reverse correlation research could not. I also demonstrate how representational similarity analysis is useful for analyzing patterns of trait ratings of aggregate classification images. This method is used frequently in fMRI research and only recently has been used by social psychologists to analyze behavioral data (Stolier et al., 2018). I am hopeful that the application of machine learning and representational similarity analysis to analyzing reverse correlation classification images will be helpful for social psychology researchers addressing a wide-range of topics.

Conclusion

The minimal group paradigm gained prominence because it showed that people discriminate even when group boundaries are meaningless. My research makes the point that we should not assume that seemingly arbitrary group distinctions are meaningless from the perspective of the people in the groups. People are motivated to find meaning in their situations and also are passively influenced by priming and spreading activation so they might latch onto associations to imbue their groups with meaning. In such situations, people may also use the self as a basis for guiding their decisions. By challenging conventional understanding of the minimal group paradigm, my work provides new insight into top-down influences of knowledge structure on face representations of minimal ingroup and outgroup.

References

Allport, G. W. (1954). *The nature of prejudice*. Reading, 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.

Andersen, S. M., & Cole, S. W. (1990). "Do I know you?": The role of significant others in general social perception. *Journal of Personality and Social Psychology, 59*(3), 384–399. https://doi.org/10.1037/0022-3514.59.3.384a

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology, 63(4)*, 596–612. https://doi.org/10.1037/0022-3514.63.4.596

Asch, S. E., & Zukier, H. (1984). Thinking about persons. *Journal of Personality and Social Psychology*, *46*(6), 1230-1240. https://doi.org/10.1037/0022-3514.46.6.1230

Ashburn-Nardo, L., Voils, C. I., & Monteith, M. J. (2001). Implicit associations as the seeds of intergroup bias: How easily do they take root? *Journal of Personality and Social Psychology, 81*(5), 789-799.

Augoustinos, M., & Rosewarne, D. L. (2001). Stereotype knowledge and prejudice in children. *British Journal of Developmental Psychology, 19*, 143-156.

Axelrod, V., Bar, M., & Rees, G. (2015). Exploring the unconscious using faces. *Trends in Cognitive Sciences, 19*, 35-45.

Bach, P., & Schenke, K. C. (2017). Predictive social perception: Towards a unifying framework from action observation to person knowledge. *Social and Personality Psychology Compass, 11,* 1-17.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10, 122–142.

Bernstein, M. J., Young, S. G., & Hugenberg, K. (2007). The cross-category effect: Mere social categorization is sufficient to elicit an own-group bias in face recognition. *Psychological Science*, *18*(8), 706-712. https://doi.org/10.1111/j.1467-9280.2007.01964.x

Billig, M., & Tajfel, H. (1973). Social categorization and similarity in intergroup behaviour. *European Journal of Social Psychology, 3*(1), 27-52. https://doi.org/10.1002/ejsp.2420030103

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science*, *15*, 674-679.

Brewer, M. B., & Silver, M. (1978). Ingroup bias as a function of task characteristics. *European Journal of Social Psychology, 8(3)*, 393-400. https://doi.org/10.1002/ejsp.2420080312

Brinkman, L., Todorov, A., & Dotsch, R. (2017). Visualising mental representations: A primer on noise-based reverse correlation in social psychology. *European Review of Social Psychology, 28*(1), 333-361.

Brown, R. J., & Abrams, D. (1986). The effects of intergroup similarity and goal interdependence on intergroup attitudes and task performance. *Journal of Experimental Social Psychology*, *22*, 78-92.

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review, 64(2)*, 123-152. https://doi.org/10.1037/h0043805

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science, 6(1)*, 3-5. doi:10.1177/1745691610393980

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42(1)*, 116-131. https://doi.org/10.1037/0022-3514.42.1.116

Cadinu, M., & Rothbart, M. (1996). Self-Anchoring and Differentiation Processes in the Minimal Group Setting. *Journal of Personality and Social Psychology, 70*. 661-77. 10.1037/0022-3514.70.4.661.

Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of Experimental Social Psychology, 55*, 110-125. https://doi.org/10.1016/j.jesp.2014.06.007

Clement, R. W., & Krueger, J. (2002). Social categorization moderates social projection. *Journal of Experimental Social Psychology, 38*(3), 219–231. https://doi.org/10.1006/jesp.2001.1503

Cogsdill, E. J., Todorov, A. T., Spelke, E. S., & Banaji, M. R. (2014). Inferring character from faces: a developmental study. *Psychological Science, 25,* 1132-1139.

Columb, C., & Plant, E. A. (2016). The Obama effect six years later: The effect of exposure to Obama on implicit anti-black evaluative bias and implicit racial stereotyping. *Social Cognition, 34*(6), 523-543.

Cone, J., Brown-Iannuzzi, J. L., Lei, R., & Dotsch, R. (2020). Type I error is inflated in the two-phase reverse correlation procedure. *Social Psychological and Personality Science,* 1-9.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. doi:10.1007/BF00994018

Crocker, J., & Luhtanen, R. (1990). Collective self-esteem and ingroup bias. *Journal of Personality and Social Psychology, 58*(1), 60–67. https://doi.org/10.1037/0022-3514.58.1.60

Crocker, J., Thompson, L. L., McGraw, K. M., & Ingerman, C. (1987). Downward comparison, prejudice, and evaluations of others: Effects of self-esteem and threat. *Journal of Personality and Social Psychology, 52*(5), 907–916. https://doi.org/10.1037/0022-3514.52.5.907

DeBruine, L. M. (2004). Resemblance to self increases the appeal of child faces to both men and women. *Evolution and Human Behavior, 25*(3), 142–154.

Den Ouden, H. E. M., Kok, P., & de Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology, 548*, 1-12.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5-18.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895-1923. doi:10.1162/089976698300017197

Dotsch, R. (2016). Rcicr: Reverse-correlation image-classification toolbox (Version 0.3.4.1). Retrieved from https://CRAN.R-project.org/package=rcicr

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science*, *3*(5), 562-571. doi:10.1177/1948550611430272

Dotsch, R., Wigboldus, D. H., Langner, O., & van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, *19*(10), 978-980. doi:10.1111/j.1467-9280.2008.02186.x

Dotsch, R., Wigboldus, D. H. J., & van Knippenberg, A. (2011). Biased allocation of faces to social categories. *Journal of Personality and Social Psychology*, *100*(6), 999-1014. doi:10.1037/a0023026

Dunham, Y., Baron, A. S., & Carey, S. (2011). Consequences of "minimal" group affiliations in children. *Child Development, 82*(3), 793-811.

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychological Science*, *17*, 383-386.

Feingold, A. (1992). Good-looking people are not what we think. *Psychological Bulletin, 111(2)*, 304-341.

Freeman, J. B., & Johnson, K. L. (2016). More than meets the eye: Split-second social perception. *Trends in Cognitive Sciences, 20,* 362-374.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*, 815–836. doi: 10.1098/rstb.2005.1622

Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? *Brain*, *123*(7), 1293-1326. doi:10.1093/brain/123.7.1293

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350–363. https://doi.org/10.1038/nrn3476

Gramzow, R. H., & Gaertner, L. (2005). Self-Esteem and Favoritism Toward Novel In-Groups: The Self as an Evaluative Base. *Journal of Personality and Social Psychology, 88(5)*, 801–815. https://doi.org/10.1037/0022-3514.88.5.801

Gramzow, R. H., Gaertner, L., & Sedikides, C. (2001). Memory for in-group and out-group information in a minimal group context: The self as an informational base. *Journal of Personality and Social Psychology*, *80*(2), 188-205.

Hassin, R., & Trope, Y. (2000). Facing faces: Studies on the cognitive aspects of physiognomy. *Journal of Personality and Social Psychology, 78*, 837-852.

Hastie, R. (1984). Causes and effects of causal attribution. *Journal of Personality and Social Psychology, 46(1)*, 44-56.

Hehman, E., Flake, J. K., & Freeman, J. B. (2018). The Faces of Group Members Share Physical Resemblance. *Personality and Social Psychology Bulletin*, *44*(1), 3–15. https://doi.org/10.1177/0146167217722556

Hogg, M. A., & Abrams, D. (1990). Social motivation, self-esteem and social identity. In D. Abrams & M. A. Hogg (Eds.), *Social identity theory: Constructive and critical advances* (pp. 28-47). New York: Harvester Wheatsheaf.

Holtz, R., & Miller, N. (1985). Assumed similarity and opinion certainty. *Journal of Personality and Social Psychology*, *48*(4), 890–898. doi:10.1037/0022-3514.48.4.890

Howard, J. M., & Rothbart, M. (1980). Social categorization and memory for in-group and out- group behavior. *Journal of Personality and Social Psychology*, *38*, 301-310.

Hugenberg, K., & Corneille, O. (2009). Holistic processing is tuned for in-group faces. *Cognitive Science, 33*, 1173-1181.

Imhoff, R., Dotsch, R., Bianchi, M., Banse, R., & Wigboldus, D. H. J. (2011b). Facing Europe: Visualizing Spontaneous In-Group Projection. *Psychological Science*, *22*(12), 1583–1590. https://doi.org/10.1177/0956797611419675

Imhoff, R., Schmidt, A. F., Bernhardt, J., Dierksmeier, A., & Banse, R. (2011a). An inkblot for sexual preference: A semantic variant of the affect misattribution procedure. *Cognition and Emotion, 25*(4), 676-690. doi:10.1080/02699931.2010.508260

Kaul, C., Ratner, K. G., & Van Bavel, J. J. (2014). Dynamic representations of race: Processing goals shape race decoding in the fusiform gyri. *Social Cognitive and Affective Neuroscience, 9(3)*, 326–332. https://doi.org/10.1093/scan/nss138

Köhler, W. (1929). *Gestalt psychology*. New York: Liveright.

Kriegeskorte, N. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(4), 1-28. doi: 10.3389/neuro.06.004.2008

Krueger, J., & Zeiger, J. S. (1993). Social categorization and the truly false consensus effect. *Journal of Personality and Social Psychology*, *65*(4), 670–680.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing." *Psychological Review, 103(2)*, 263-283. https://doi.org/10.1037/0033-295X.103.2.263

Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and

application. *Psychological Bulletin, 129*(4), 522-544. https://doi.org/10.1037/0033-2909.129.4.522

Lazerus, T., Ingbretsen, Z. A., Stolier, R. M., Freeman, J. B., & Cikara, M. (2016). Positivity bias in judging ingroup members' emotional expressions. *Emotion, 16*, 1117-1125.

Levari, D. E., Gilbert, D. T., Wilson, T. D., Sievers, B., Amodio, D. M., & Wheatley, T. (2018). Prevalence-induced concept change in human judgment. *Science*, *360*(6396), 1465-1467.

Lindström, B., Selbing, I., Molapour, T., & Olsson, A. (2014). Racial bias shapes social reinforcement learning. *Psychological Science, 25*(3), 711-719.

Locksley, A., Ortiz, V., & Hepburn, C. (1980). Social categorization and discriminatory behavior: Extinguishing the minimal intergroup discrimination effect. *Journal of Personality and Social Psychology, 39*(5), 773-783.

Luhtanen, R., & Crocker, J. (1992). A collective self-esteem scale: Self-evaluation of one's social identity. *Personality and Social Psychology Bulletin, 18*(3), 302–318. https://doi.org/10.1177/0146167292183006

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska directed emotional faces - KDEF. CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet. Retrieved from http://kdef.se/home/aboutKDEF.html

Maclnnis, C. C., & Page-Gould, E. (2015). How can intergroup interaction be bad if intergroup contact is good? Exploring and reconciling an apparent paradox in the science of intergroup relations. *Perspectives on Psychological Science, 10*, 307-327.

Macrae, C. N., Quinn, K. A., Mason, M. F., & Quadflieg, S. (2005). Understanding others: The face and person construal. *Journal of Personality and Social Psychology, 89*, 686-695.

Maddox, K. B. (2004). Perspectives on racial phenotypicality bias. *Personality and Social Psychology Review, 8(4)*, 383-401. https://doi.org/10.1207/s15327957pspr0804_4

Mangini, M., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science*, *28*(2), 209-226. doi:10.1016/j.cogsci.2003.11.004

Mattarozzi, K., Todorov, A., Marzocchi, M., Vicari, A., & Russo, P. M. (2015). Effects of Gender and Personality on First Impression. *PLOS ONE*, *10*(9), e0135529. https://doi.org/10.1371/journal.pone.0135529

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44(12)*, 1469-1481. http://dx.doi.org/10.1037/0003-066X.44.12.1469

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2019). E1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), TU wien (Version 1.7-1). Retrieved from https://CRAN.R-project.org/package=e1071

Murphy, G. L., & Medin, D. L. (1985). The Role of Theories in Conceptual Coherence. *Psychological Review, 92(3)*, 289-316. https://doi.org/10.1037/0033-295X.92.3.289

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231-259. doi:10.1037/0033-295X.84.3.231

Ojala, M., & Garriga, G. C. (2009). Permutation tests for studying classifier performance. In *2009 ninth IEEE international conference on data mining* (pp. 908-913). Miami Beach, FL, USA: IEEE. doi:10.1109/ICDM.2009.108

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087-11092. doi:10.1073/pnas.0805664105

Otten, S., & Moskowitz, G. B. (2000). Evidence for implicit evaluative in-group bias: Affect-biased spontaneous trait inference in a minimal group paradigm. *Journal of Experimental Social Psychology, 36*(1), 77-89.

Park, B., & Judd, C. M. (1990). Measures and models of perceived group variability. *Journal of Personality and Social Psychology, 59(2)*, 173-191.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*(3), 277-293. https://doi.org/10.1037/0022-3514.89.3.277

Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology, 49*, 65-85.

Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta- analytic tests of three mediators. *European Journal of Social Psychology, 38*, 922-934.

Pinter, B., & Greenwald, A. G. (2010). A comparison of minimal group induction procedures. *Group Processes & Intergroup Relations, 14*(1), 81-98.

Qian, M. K., Heyman, G. D., Quinn, P. C., Fu, G., & Lee, K. (2017). When the majority becomes the minority: A longitudinal study of the effects of immersive experience

with racial out-group members on implicit and explicit racial biases. *Journal of Cross-Cultural Psychology, 48*(6), 914-930.

Rabbie, J. M., & Horwitz, M. (1969). Arousal of ingroup-outgroup bias by a chance win or loss. *Journal of Personality and Social Psychology, 13(3)*, 269-277. https://doi.org/10.1037/h0028284

Ratner, K. G., & Amodio, D. M. (2013). Seeing "us vs. them": Minimal group effects on the neural encoding of faces. *Journal of Experimental Social Psychology, 49*(2), 298-301.

Ratner, K. G., Dotsch, R., Wigboldus, D. H. J., van Knippenberg, A., & Amodio, D. M. (2014). Visualizing minimal ingroup and outgroup faces: Implications for impressions, attitudes, and behavior. *Journal of Personality and Social Psychology*, *106*(6), 897-911.

Rosch, E., & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573-605. https://doi.org/10.1016/0010-0285(75)90024-9

Rosenberg, M. (1965). Society and the Adolescent Self-Image. Princeton, NJ: Princeton University Press.

Rothbart, M., & Taylor, M. (1992). Category labels and social reality: Do we view social categories as natural kinds? *In Language, interaction and social cognition* (pp. 11-36). Thousand Oaks, CA, US: Sage Publications, Inc.

Scholkopf, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T., & Vapnik, V. (1997). Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, *45*(11), 2758-2765.

Sherman, J. W., Stoessner, S. J., Conrey, F. R., & Azam, O. A. (2005). Prejudice and stereotype maintenance processes: attention, attribution, and individuation. *Journal of Personality and Social Psychology, 89*, 607-622.

Sidhu, D. M., Deschamps, K., Bourdage, J. S., & Pexman, P. M. (2019). Does the name say it all? Investigating phoneme-personality sound symbolism in first names. *Journal of Experimental Psychology: General, 148*(9), 1595-1614. http://dx.doi.org/10.1037/xge0000662

Srull, T. K., & Wyer, R. S. (1989). Person memory and judgment. *Psychological Review, 96,* 58-83.

Stolier, R. M., & Freeman, J. B. (2016). Neural pattern similarity reveals the inherent intersection of social categories. *Nature Neuroscience*, *19*(6), 795-797. doi:10.1038/nn.4296

Stolier, R. M., Hehman, E., & Freeman, J. B. (2018). A dynamic structure of social trait space. *Trends in Cognitive Sciences*, *22*(3), 197-200. doi:10.1016/j.tics.2017.12.003

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149-178. doi:10.1002/ejsp.2420010202

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin, & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33-37). Monterey, CA: Brooks/Cole.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103(2)*, 193-210.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 10,* 1623-1626.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12*, 455-460.

Van Bavel, J. J., & Cunningham, W. A. (2009). Self-categorization with a novel mixed-race group moderates automatic social and racial biases. *Personality and Social Psychology Bulletin*, *35*(3), 321-335. https://doi.org/10.1177/0146167208327743

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2008). The neural substrates of in-group bias: A functional magnetic resonance imaging investigation. *Psychological Science, 19*(10), 1131-1139. doi: 10.1111/j.1467-9280.2008.02214.x

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2011). Modulation of the fusiform face area following minimal exposure to motivationally relevant faces: Evidence of in-group enhancement (not out-group disregard). *Journal of Cognitive Neuroscience, 23*(11), 3343- 3354. doi: 10.1162/jocn_a_00016

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review, 101*(1), 34-52. https://doi.org/10.1037/0033-295X.101.1.34

Weibert, K., Flack, T. R., Young, A. W., & Andrews, T. J. (2018). Patterns of neural response in face regions are predicted by low-level image properties. *Cortex*, *103*, 199–210. https://doi.org/10.1016/j.cortex.2018.03.009

Weisbuch, M., Pauker, K., Adams, R. B., Jr., Lamer, S. A., & Ambady, N. (2017). Race, power, and reflexive gaze following. *Social Cognition, 35*(6), 619-638.

Wilson, J. P., Young, S. G., Rule, N. O., & Hugenberg, K. (2018). Configural processing and social judgments: Face inversion particularly disrupts inferences of human-relevant traits. *Journal of Experimental Social Psychology, 74*, 1-7.

Young, S. G., & Hugenberg, K. (2010). Mere social categorization modulates identification of facial expressions of emotion. *Journal of Personality and Social Psychology, 99*, 964– 977.