Journal of Personality and Social Psychology

Minimal but Not Meaningless: Seemingly Arbitrary Category Labels Can Imply More Than Group Membership

Youngki Hong and Kyle G. Ratner

Minimal but Not Meaningless: Seemingly Arbitrary Category Labels Can Imply More Than Group Membership

Youngki Hong and Kyle G. Ratner
University of California, Santa Barbara

Minimal group paradigms tend to involve contrived group distinctions, such as dot estimation tendencies and aesthetic preferences. Researchers assume that these novel category distinctions lack informational value. Our research tests this notion. Specifically, we used the classic *overestimator versus underestimator* and *Klee versus Kandinsky* minimal group paradigms to assess how category labels influence minimal group responses. In Study 1, we show that participants represented ingroup faces more favorably than outgroup faces, but also represented overestimator and underestimator category labels differently. In fact, the category label effect was larger than the intergroup effect, even though participants were told that estimation tendencies were unrelated to other cognitive tendencies or personality traits. In Study 2, we demonstrate that Klee and Kandinsky were also represented differently, but in this case, the intergroup effect was stronger than the category label effect. In Studies 3 and 4, we examined effects of category labels on how participants allocate resources to, evaluate, and ascribe traits to ingroup and outgroup members. We found both category label and intergroup effects when participants were assigned to overestimator and underestimator groups. However, we found only the intergroup effect when participants were assigned to Klee and Kandinsky groups. Together, this work advances but does not upend understanding of minimal group effects. We robustly replicate minimal intergroup bias in mental representations of faces, evaluations, trait inferences, and resource allocations. At the same time, we show that seemingly arbitrary category labels can imply characteristics about groups that may influence responses in intergroup contexts.

Keywords: minimal group paradigm, reverse correlation, machine learning, representational similarity analysis, resource allocation

One only has to be a casual reader of social psychology to know about the minimal group paradigm and the dogma that merely separating people into arbitrary groups creates a variety of intergroup biases. However, are the group distinctions used in this research really arbitrary? It is no coincidence that Tajfel, Billig, Bundy, and Flament (1971) used rather contrived group distinctions, *overestimators versus underestimators* and preference for paintings by *Klee versus Kandinsky*, in their landmark article. They were aware that 2 years prior, Rabbie and Horwitz (1969) reported that when group distinctions were completely arbitrary then impressions of novel ingroup and outgroup members did not differ. In reference to Rabbie and Horwitz's research, Tajfel et al. (1971) demurred that expecting purely random groupings to produce intergroup bias would be as nonsensical as expecting people

to show biases based on "sitting on the same and opposite benches in a compartment of a train" (p. 152).

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. 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, Young, & Hugenberg, 2007) or common fate implied by a competitive context (e.g., Cikara, Bruneau, Van Bavel, & Saxe, 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.

Among researchers, the belief has always been that the category labels might differ but these differences do not carry meaning in the intergroup context beyond demarcating ingroup and outgroup. Tajfel (1970) set the stage for this long-held assumption by calling the category labels "artificial and insignificant" (p. 97) and "flimsy and unimportant criteria" (p. 101). This belief is so entrenched that many researchers do not even report whether they carefully counterbalanced the category labels or provide statistics showing that category labels do not matter. However, it is merely conjecture that the typical minimal group labels are stripped of their inferential qualities. Do people really see overestimators and underestimators as the same? What about Klee and Kandinsky fans? Is it possible that people read

into the meaning of novel category labels and infer underlying attributes in ways that influence intergroup responses? The conventional wisdom is that labels do not matter, but if they do, then this unacknowledged truth has been hiding in plain sight since Tajfel's foundational research.

Why Category Labels Might Matter

In a minimal group situation, participants are confronted with categories that are plausible but novel to them. The novelty of the categories 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, Pauker, Adams, Lamer, & Ambady, 2017), stereotypes circulating in the culture (Devine, 1989), personal antipathy (Augoustinos & Rosewarne, 2001), direct experience with the groups (Columb & Plant, 2016; Qian, Heyman, Quinn, Fu, & Lee, 2017), own group preference (Lindström, Selbing, Molapour, & Olsson, 2014), or other confounded variables. Conventional wisdom in social psychology is that minimal groups sidestep this problem because novel categories by virtue of their novelty signal whether a target shares one's group membership or not, but do not convey much else. However, just because this is what experimenters want and assume to be true does not mean that participants will allow it to be so.

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. How should they go about solving the task at hand? With established groups such as race, perceivers are usually able to draw on their knowledge of stereotypes and life experiences with members of the groups to guide their judgments and behaviors (Kunda & Spencer, 2003). Without this concrete knowledge, it is assumed by researchers that participants will default to a heuristic that ingroups should be preferred. For example, social identity theory suggests that this occurs to maintain self-esteem (Hogg & Abrams, 1990; Tajfel & Turner, 1979). Evolutionary perspectives argue that ingroups should be preferred because ingroup members are the conspecifics who through phylogenetic history have provided coalitional support to ward off threats and other support that allowed individuals to thrive (Cosmides, Tooby, & Kurzban, 2003; Miller, Maner, & Becker, 2010). These perspectives see the category label as a mechanism to signal who is an ingroup member and who is an outgroup member and once it has served this purpose it is discarded like a fuel tank that has launched a rocket into orbit. 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. This overt emphasis on the category labels gives participants several reasons to not conform to the experimenter's desires and instead try to make sense of minimal group labels to guide their task decisions.

First, if one looks at the minimal group situation as a communicative context between the experimenter and perceiver, several of Grice's Maxims are applicable to understand the pragmatics of this situation (Blank, 1997, 2004; Grice, 1975). Grice's Maxim of Quality says that people tend to make statements that are truthful

and supported by facts, so the participants should default to believing the minimal group distinctions that the experimenter asserts, even if they seem convoluted. Furthermore, Grice's Maxim of Relation suggests that people only share relevant information, so participants might assume that there must be a deeper meaning to the category labels, otherwise the experimenter would not take the time to study them. However, even if participants trust the claims about the category labels and assume that the distinctions must be important, how would participants go about inferring meaning from category labels that are designed to be meaning poor?

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 we 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 situation, the category labels might provide information to 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. The label overestimator is objectively defined 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. Maybe the halo of being optimistic or dominant leads people to value overestimators more than underestimators. Thus, perceivers can very quickly go beyond the information given. To be clear, we are not arguing that inferences about the meaning of category labels supplant intergroup bias, but it is possible that they have an unrecognized effect on responding.

Some Category Labels Might Provide More to Latch Onto Than Others

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 explained above, it is rather clear to see how overestimators and underestimators might be 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 is the case with 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. 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.

Some Tasks Might Provide More Reasons to Latch Onto Category Labels Than Others

Over the years, minimal group effects have been shown in a myriad of domains, ranging from resource allocations (Tajfel et al., 1971), to explicit attitudes and trait inferences (Brewer & Silver, 1978; Dunham, Baron, & Carey, 2011; Otten & Moskowitz, 2000), to memory for person information (Bernstein et al., 2007; Gramzow, Gaertner, & Sedikides, 2001), to implicit attitudes (Ashburn-Nardo, Voils, & Monteith, 2001; Dunham et al., 2011), to face representation (Ratner, Dotsch, Wigboldus, van Knippenberg, & Amodio, 2014), and face perception (Hugenberg & Corneille, 2009; Ratner & Amodio, 2013; Van Bavel, Packer, & Cunningham, 2008, 2011). Across all of these domains, a rather consistent pattern of ingroup preference is observed. Yet, given the differing task demands that are necessary for these various types of processing, it might be the case that category labels have more of an influence on some of these processing modalities than others.

Take for instance, a recent demonstration by Ratner, Dotsch, Wigboldus, van Knippenberg, and Amodio (2014) that people visualize minimal ingroup faces differently than minimal outgroup faces. The task demands of visualization are particularly onerous. It is difficult to visualize an abstract distinction, such as ingroup versus outgroup. The reason for this is that visualization is most vivid when concrete details are available, which at a minimum occurs at the basic level of categorization (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). This is where the category labels could become meaningful. As mentioned earlier, several successful instantiations of the minimal group paradigm use labels that allow deeper qualities of the people who fit into the group to be inferred, including the overestimator and underestimator distinction used by Ratner et al. (2014). Thus, instead of relying on an abstract distinction like ingroup and outgroup to infer qualities of a face, it might be easier for participants to imagine what an overestimator or underestimator should look like and this more

concrete representation could then provide a basis from which ingroup bias is demonstrated.

Compare this with the task demands of making resource allocation decisions—the main dependent variable in Tajfel et al.'s (1971) classic studies. Deciding whether to allocate resources to another individual does not require a detailed visualization of their face. Thus, if one's goal is to allocate resources without any information besides a novel category label, then it might be cognitively efficient to not worry about the meaning of the minimal group label and use the heuristic that ingroups should be favored to guide one's decisions. For these reasons, the extent to which category labels might have an effect in a minimal group situation could depend on the processing goals of the perceiver (e.g., visualizing faces vs. allocating resources).

Overview of the Studies

The current set of studies were designed to examine two primary objectives. The first was to determine if people imbue classic minimal group labels with any meaning whatsoever. The second was to examine whether the inductive potential of the labels and the processing goals of the perceiver are critical levers on whether category labels have an influence on responses in an intergroup context. To address this first aim we 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, Todorov, & Dotsch, 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. Study 1a uses a preregistered, highly powered replication of Ratner et al.'s (2014) Study 1 to first establish that their intergroup bias findings were not simply false positives. It goes beyond Ratner et al. (2014), however, by also examining representational differences between overestimator and underestimator at the group-level. 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. Study 1b uses representational similarity analysis (Kriegeskorte, Mur, & Bandettini, 2008; Stolier & Freeman, 2016; 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. Study 1c uses a machine learning approach to examine the representational differences between overestimator and underestimator at the participant-level to provide convergent support for the results identified in Study 1a. Study 2 examines the generalizability of these phenomena by replicating the findings from Study 1 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). Given that the gold standard for consequential behavioral effects of minimal group paradigms is resource allocations, Studies 3 and 4 examine whether any representational effect of overestimator and underestimator and Klee and Kandinsky have meaningful behavioral implications on a resource allocation task. Overt trait ascriptions and evaluations were also measured. By using minimal group paradigms that vary to the extent that attributes can be logically inferred and also tasks that vary in how

functional it might be to use category label information, these studies collectively test the assumption that Tajfel and many researchers who followed him made about minimal group labels—that they are flimsy and uninformative.

Study 1

There were multiple goals of Study 1. First, we attempted to replicate the findings of Ratner et al. (2014) by demonstrating that people show ingroup positivity in face representations. Second, we tested for differences in people's face representations of overestimators and underestimators regardless of whether they are ingroup or outgroup. Third, after establishing that minimal group labels might have meaningful distinctions in face representations, we sought to understand how the ingroup/outgroup distinction and the overestimator/underestimator distinction differentially contribute to face representations. Lastly, we 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 our data, we break down Study 1 into three parts (a, b, and c).

Study 1a

Study 1a was conducted in two phases, in which we (a) created visual renderings of participants' mental 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. We used the reverse correlation image classification technique to create visual 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, we assessed whether these visual renderings could reveal differences in face representations of different groups by asking an independent sample 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.

This study 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. We predicted that people would show ingroup favoritism as indicated by more positive trait ratings of ingroup faces than outgroup faces. We remained unsure a priori about whether the minimal group labels would have an influence on face representations.

Method

Phase 1: Generating visual renderings of face representations.

Participants. We 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. We sought to maximize our power by (a) preregistering our sample size on the Open Science Framework (https://osf.io/s9243/?view_only= 92afae84a38548e8a9412e8353f30905) and (b) 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). Our sample was obtained from the UCSB Psychological and Brain Sciences subject pool, which consisted of people from diverse backgrounds including but not limited to different genders, racial and ethnic backgrounds, religions, national origins, and political beliefs. 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. Participants provided written informed consent approved by the UCSB Human Subjects Committee.

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, we told our participants that people vary in numerical estimation style, which was defined as the tendency to overestimate or underestimate the number of objects they encounter. We 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] We then told our 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). We also told them that people can reliably detect

---

[1] This instruction, which was also provided in Ratner et al. (2014), was designed to constrain participants' impulse to read into the meaning of the category labels. However, given that people are motivated to make sense of their environment, participants might read into the meaning of the category labels anyway. Relevant to this possibility, research on transference, which is another social cognitive example of "going beyond the information given," shows that it is very difficult for participants to not show transference even when explicitly told to avoid doing so (Przybylinski & Andersen, 2013). Research on affective and semantic misattribution also shows that people fail to ignore information that they are told is irrelevant to their judgments (Imhoff et al., 2011; Payne et al., 2005).

numerical estimation style from faces and that the purpose of the current 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. We did not actually take participants' NEST responses into account; the NEST was used to provide a rationale for the group assignment.

We 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.

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

Face representation data processing. Following the logic of reverse correlation analysis, we generated visual renderings of different groups by averaging noise patterns of selected faces (Dotsch & Todorov, 2012; Dotsch, Wigboldus, Langner, & van Knippenberg, 2008; Dotsch, Wigboldus, & van Knippenberg, 2011). We 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 us to examine 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. We 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, we created eight group-level classification images. First, to test whether our findings replicate the ingroup positivity effect found in Study 1 of Ratner et al. (2014), we created ingroup ($n = 180$) and outgroup ($n = 182$) classification images by averaging the appropriate noise patterns from the participant-level. That is, we averaged noise patterns of participant-level classification images of ingroup faces and superimposed the normalized average noise pattern back onto the base image to create the group-level classification image for the ingroup face. We did the same for the outgroup face (see Figure 1). Second, to examine the difference in trait impressions elicited by the category labels, we 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, we 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.

Phase 2: Assessing impressions of face representations. In Phase 2, we objectively assessed the differences in these face representations, specifically in how they elicited different trait impressions. To do this, we 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 × NEST images, we obtained ratings from three different samples of participants. That is, participants only rated ingroup and outgroup images, overestimator and underestimator images, or Group × NEST images.

Participants. We 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. We recruited a comparable number of Mechan-
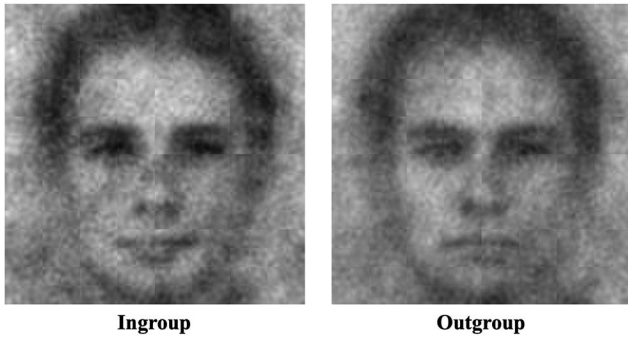
Figure 1. Study 1 ingroup and outgroup group-level classification images.

ical Turk (MTurk) raters as Ratner et al. (2014). The racial and ethnic breakdown of our sample of 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. Participants provided written informed consent approved by the UCSB Human Subjects Committee.

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.[2] 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, we conducted a repeated-measures multivariate analysis of variance (rMANOVA) followed by a univariate analysis of variance for each trait. We show the results below separated by sample.
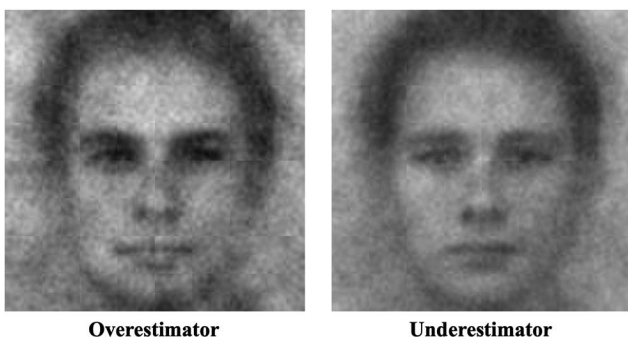

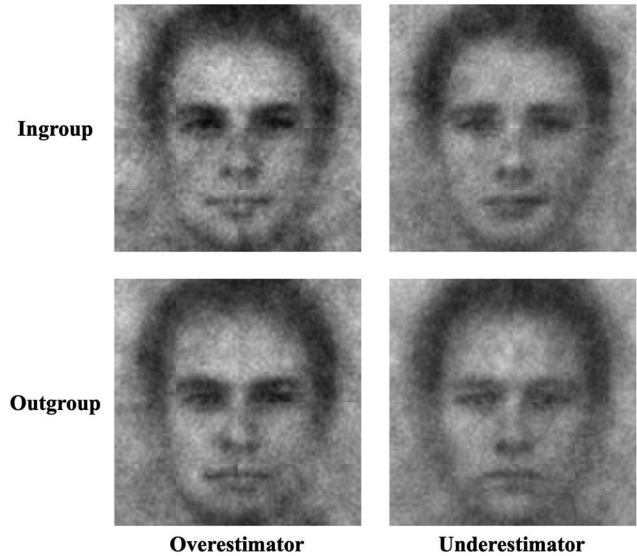**Figure 2.** Study 1 overestimator and underestimator group-level classification images.

Figure 3. Study 1 Group $\times$ NEST group-level classification images.

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.

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.

Group $\times$ NEST. We 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: Group

---

[2] This is a slight deviation from Ratner et al.'s (2014) protocol. They presented the ingroup and outgroup classification images together on the screen during the ratings, so participants could make relative comparisons when making their judgments. Our decision to present each face by itself is a stronger test of whether the classification images differ.

Table 1

Study 1 Trait Rating ANOVA Results—Group—Face Representations

Trait	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

* $p < .05$.   ** $p < .01$.   *** $p < .001$.

(Pillai's Trace = .58, $F = 30.57$, $df = (13, 285)$, $p < .001$), NEST (Pillai's Trace = .48, $F = 20.49$, $df = (13, 285)$, $p < .001$), and Group × 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.

Discussion

In Study 1a, we investigated the face representations of minimally defined groups. First, our results replicated the findings of Ratner et al. (2014): Ingroup faces elicited overall more positive trait impressions compared with outgroup faces. The current study was preregistered (https://osf.io/s9243/?view_only=92afae84a385 48e8a9412e8353f30905) and highly powered—twice the sample size of the original study by Ratner et al. (2014), providing strong evidence that their demonstration of ingroup positivity in face representations is a replicable effect.

More interestingly, however, we also found that participants generated different face representations of overestimators and underestimators. 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 differences between minimal group labels are arbitrary (Tajfel et al., 1971). Instead, findings from the current study showed that people might utilize novel category labels when visualizing faces of ingroup and outgroup members, and infer different traits from those labels. We

also found several interaction effects for the Group × 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 additional evidence that face representations of overestimator and underestimator are different and can influence intergroup bias. For instance, overestimators were generally rated as more attractive, emotionally stable, sociable, aggressive, and mean than underestimators, and perhaps this constrained variability on these trait dimensions for overestimators, which resulted in stronger ingroup and outgroup differences on these variables for underestimators. Interestingly, for the caring dimension, 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 versus underestimator distinction.

Study 1b

In Study 1b, we used multiple regression representational similarity analysis (RSA; Kriegeskorte et al., 2008; Stolier & Freeman, 2016; Stolier et al., 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 × 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. Our theoretical premise was that participants completed the face categorization task from Study 1a with two pieces of information: (a) the category label (overestimator or underestima-

Table 2

Study 1 Trait Rating ANOVA Results—NEST—Face Representations

Trait	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

Note. NEST = Numerical Estimation Style Test.
+ $p < .10$.   * $p < .05$.   ** $p < .01$.   *** $p < .001$.

Table 3
Study 1 Trait Rating ANOVA Results—Group × NEST—Face Representations

Trait	Overestimator			Underestimator			F values		
	Ingroup (SD)	Outgroup (SD)	Cohen's d	Ingroup (SD)	Outgroup (SD)	Cohen's d	Group	NEST	Group × 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**

Note. NEST = Numerical Estimation Style Test.
+ $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

tor) of the targets (NEST); and (b) 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 × NEST images should reflect linear combinations of trait representations of Group images and those of NEST images. Thus, by using multiple regression RSA, we 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 × NEST groups (ingroup-overestimator, outgroup-overestimator, ingroup-underestimator, and outgroup-underestimator) during the face categorization task.

Method

Participants. The data from the same 301 participants recruited in Phase 2 of Study 1a were reanalyzed here. As stated in Phase 2 of Study 1a, 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 Study 1a for a more detailed description.

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

regression RSA to predict correlation vectors of Group × NEST trait rating data with linear combinations of correlation vectors of appropriate Group and NEST trait rating data. For example, we 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 the overestimator group-level classification image (see Figure 4). By using multiple regression RSA, we tested unique contributions of group membership (Group) and minimal group labels (NEST) to Group × 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 participants used the ingroup/outgroup distinction more than the underestimator label when choosing ingroup underestimator faces during the face categorization task, the trait representation of ingroup should have a higher beta value than the trait representation of underestimator.

Results

Ingroup overestimator. We used ordinary least squares multiple regression to predict the pairwise correlation vector of the trait rating data of the ingroup overestimator face image with the linear combination of the correlation vectors of the ingroup face trait rating data and the overestimator face trait rating data. We 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 the ingroup overestimator ratings. We 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. We 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
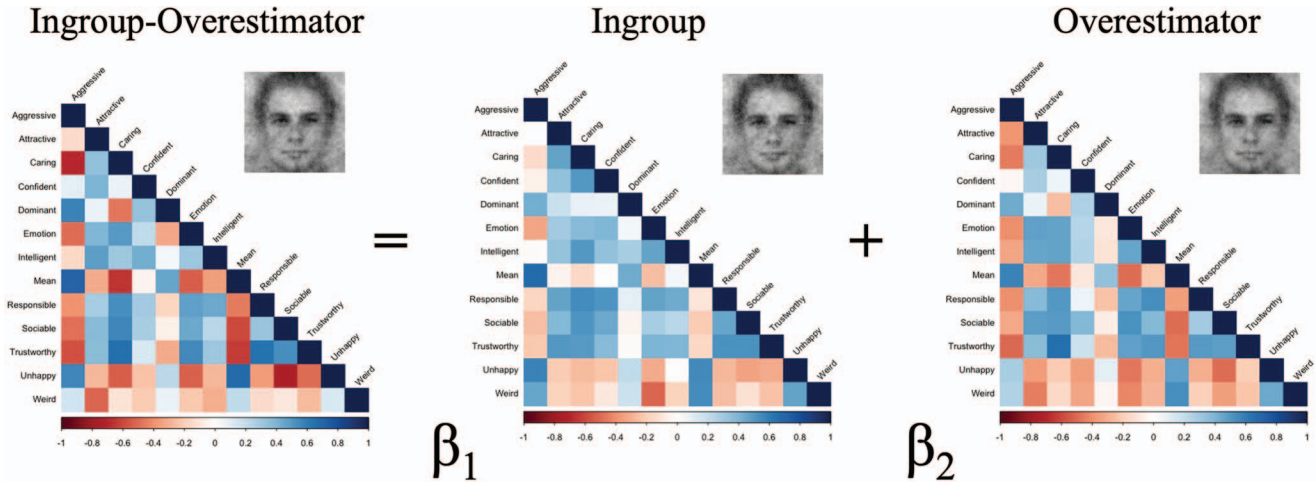
Figure 4. Multiple regression RSA example: Predicting the pairwise correlation matrix of ingroup overestimator trait rating data from the linear combination of the pairwise correlation matrices of ingroup trait rating data and overestimator trait rating data. Each square represents a pairwise correlation value. See the online article for the color version of this figure.

overestimator face trait rating data. We 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. We used multiple regression to predict ingroup underestimator trait rating data with the linear combination of ingroup face trait rating data and underestimator face trait rating data. We 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. We used multiple regression to predict outgroup underestimator trait rating data with the linear combination of outgroup face trait rating data and underestimator face trait rating data. We 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

Study 1b examined how people generated face representations of different minimal groups using multiple regression RSA on trait rating data of group-level classification images. We found that for both ingroup overestimator and outgroup overestimator face representations, participants seemed to have used the overestimator label more than the ingroup/outgroup distinction, as indicated by larger trait

representational similarities between Group × 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 × 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, as indicating 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 with 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 our interpretations of the 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), we examined whether the difference we found in mental representations of

overestimator and underestimator faces holds at the participant level in Study 1c.

Study 1c

Study 1a and 1b provided evidence that face representations of minimally defined groups can vary and lead to trait impressions that differ on various dimensions of social perception and that different minimal group labels have different degrees 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 Study 1c we tested whether representational differences found in trait impressions of different minimal groups in Study 1a also exist at the participant 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, we 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.

Method

Stimuli. In Study 1c, we used 362 participant-level classification images from Phase 1 of Study 1a. Each image had three dimensions: (a) Group (ingroup or outgroup); (b) NEST (overestimator or underestimator); and (c) Group × NEST (ingroup-overestimator, outgroup-overestimator, ingroup-underestimator, or outgroup-underestimator).

Procedure. We used the R package *e1071* (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2019) to conduct the machine learning analyses following three steps: (a) vectorizing and down sampling pixel intensity data of each image (see Figure 5); (b) standard scaling (i.e., standardization); and (c) classification using support vector machines (SVM) with a radial basis function (RBF) kernel (Cortes & Vapnik, 1995; Scholkopf et al., 1997). We performed cross-validation for each analysis to ensure that every image was in (not at the same time) both training and testing data sets. We down sampled participant-level classification images by simply resizing them from 512×512 pixels to 64×64 pixels. We conducted the same true label analysis using 512×512, 256×256, 128×128, and 64×64 image sizes, and found no detrimental effect of down sampling on the classification accuracies. Thus, we downsampled the images due to the computationally
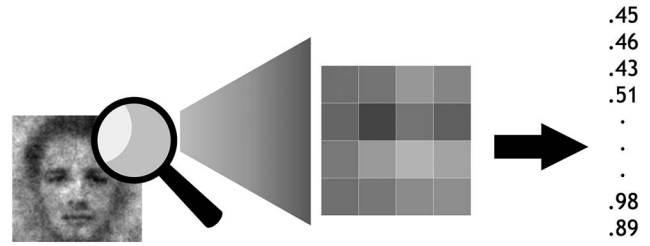

Figure 5. Example of vectorizing pixel intensity data from a participant-level classification image.

intensive nature of machine learning analyses needed for the 1,000 permutation tests. Standard scaling was done by mean-centering pixel intensity data and dividing the values by their standard deviation. Finally, we 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 = 64×64). We used this same procedure to classify between ingroup and outgroup faces, overestimator and underestimator faces, and Group × NEST faces.

To minimize overfitting and maximize our chance of detecting real differences in classification images, we used 10-fold cross-validation with our SVM model. 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 (64×64 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. We repeated this step 10 times until every instance of data was in both training and testing sets at some point. We then computed accuracy scores by averaging accuracies from these 10 folds. By using this method of cross-validation, we 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, we used permutation tests to determine whether accuracies of our SVM classifications differed significantly from chance (Ojala & Garriga, 2010). For each permutation, class labels (e.g., ingroup or outgroup) were randomly permuted for every image, followed by the classification steps described above. We repeated the same procedure 1,000 times, creating our own null distribution against which we could compare the accuracies of our classification results with true labels. We 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).

We also compared the accuracies of our model's classifications of Group and NEST labels using the 5 × 2-fold cross-validation paired samples *t* test (Dietterich, 1998). That is, we performed five replications of twofold cross-validation (splitting data into equal number of training and testing data), resulting in 10 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. We 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 nine 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 5×2-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.

Results

We were able to classify between ingroup and outgroup images from pixel intensity data significantly better than chance (accuracy = 59.20%, $p < .001$). The same was true for overestimator and underestimator images (accuracy = 66.01%, $p < .001$). For both classifications, all permutation tests yielded lower accuracy scores than the accuracy scores with true labels (see Figure 6). Next, we compared classification accuracies for Group and NEST using the 5×2-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 our 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 [3.66, 10.60].

Finally, multiclass SVM results showed that our model performed significantly better than chance (accuracy = 37.83%, $p < .001$). Unlike the previous two cases, the chance accuracy for the current classification was 25% (one out of four). Upon examining the confusion matrix of results using the true labels, we found that our 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 Study 1c, we 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. We examined the differences between Group (ingroup and outgroup), NEST (overestimator and underestimator), and Group $\times$ NEST participant-level classification images using a novel approach for analyzing reverse correlation images, specifically a machine learning algorithm called support vector machine. We found that using this method we could classify Group, NEST, and Group $\times$ NEST participant-level classification images better than chance, suggesting that the differences between face representations of all category types exist at the participant level.

We also found that the SVM classified between overestimator and underestimator face images significantly better than ingroup and outgroup face images, providing a piece of evidence that NEST labels were used more than the ingroup/outgroup distinction 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 our face categorization task created 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 Study 1b may partly address this possibility. Specifically, the more "consistent" face representation of overestimator and underestimator versus ingroup and outgroup found in Study 1c was true mostly for overestimator faces but not necessarily for underestimator faces in Study 1b. Thus, we argue that minimal group labels can be meaningful, albeit to different extents for different labels.

In short, we 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. We also did not use subjective trait ratings to arrive at this conclusion, thus providing stronger evidence that ingroup and outgroup faces as well as overestimator and underestimator faces are objectively different from each other. We were also able to show the same findings as Study 1a despite the fact that we used very different methods (i.e., trait ratings vs. image classification using pixel intensity data), suggesting that representational biases that arise with this minimal group paradigm are robust.

The finding of more misclassifications within NEST labels than within Group labels of Group $\times$ 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 minimal group labels seemed to
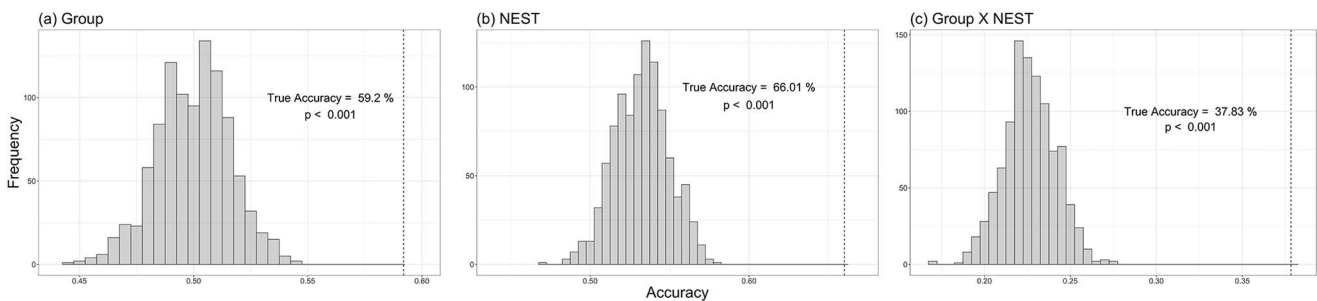

Figure 6. Permutation test results of (a) Group, (b) NEST, and (c) Group $\times$ NEST. The dotted lines indicate true accuracy scores.

have played a greater role in the face categorization task than whether the targets shared the same group membership with the participant or not.

Study 2

So far, we showed that people have different mental representations of 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 like. Thus, people seem to imbue meaning to minimal group labels but to different extents for different minimal groups. One critical limitation is that we used only one version of the minimal group paradigm (i.e., the overestimator vs. underestimator distinction), therefore we have not shown whether people imbue meaning to other minimal group labels (e.g., Klee vs. Kandinsky). Additionally, although we 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 minimal group paradigms to investigate other forms of intergroup bias. The study to follow investigated the generalizability of our findings from Study 1 with a different type of minimal group paradigm, the Klee versus Kandinsky distinction (Experiment 2 from Tajfel et al., 1971). As discussed previously, the Klee and Kandinsky paradigm differs from the overestimator and underestimator paradigm in that the former on its face seems to have labels with less inductive potential than does the latter. Study 2 used the same set of methods from Study 1 to empirically examine whether people represent faces of people who like Klee paintings differently from faces of those who like Kandinsky paintings.

Study 2a

Following the procedure of Study 1a, Study 2a was also conducted in two phases. In Phase 1, participants were randomly assigned to minimal groups (Klee vs. Kandinsky groups) and then categorized faces as belonging to either of these two minimal groups. We used the reverse correlation image classification technique to create visual representations of Klee and Kandinsky fans as well as ingroup and outgroup faces. In Phase 2, we assessed whether images of these different minimal group faces would be rated differently by independent samples of participants on the 13 trait dimensions used in Study 1.

Although we found some differences in trait impressions between different minimal groups (overestimator vs. underestimator) from Study 1, we 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 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), we 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. We recruited 200 University of California, Santa Barbara students (M_{age} = 18.82, SD = 1.07; 149 female, 47 male,

and four unidentified) to participate in a study about categorizing faces in exchange for course credit. We preregistered our sample size on the Open Science Framework (https://osf.io/s9243/?view_only=92afae84a38548e8a9412e8353f30905). Our sample was from the UCSB Psychological and Brain Sciences subject pool, which consisted of people from diverse backgrounds, including but not limited to different genders, racial and ethnic backgrounds, religions, national origins, and political beliefs. The racial and ethnic breakdown of our sample was 65 Asian, 65 White, 35 Latinx, 24 multiracial, five other, and six unidentified participants. Up to four participants were run simultaneously. Participants provided written informed consent approved by the UCSB Human Subjects Committee.

Procedure. The current study followed the same 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, we used a classic *aesthetic preference* procedure (Experiment 2 from Tajfel et al., 1971) to assign participants to novel, but believable, groups. Then they conducted a face categorization task optimized for a reverse correlation analysis.

Artistic Preference Test (ART). In this task, we told our participants that people can reliably figure out another person's artistic preference simply by looking at their face. We then told our participants that they would categorize photographs of students from a previous quarter whose artistic preference had been determined. We 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 (counterbalanced across participants), indicating that each participant had a preference for paintings by either Kandinsky or Klee. We did not actually take participants responses into account; the ART was used to provide a rationale for the group assignment.

We used additional procedures to make the novel groups (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 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 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 gray-
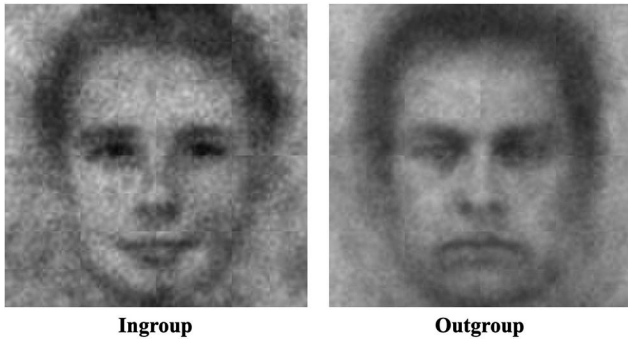
Figure 7. Study 2 ingroup and outgroup group-level classification images.

Figure 8. Study 2 Klee and Kandinsky group-level classification images.

scale 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. We used the same set of face stimuli from Study 1a—450 pairs of face images generated from the grayscale neutral male average face of the Averaged Karolinska Directed Emotional Faces Database (Lundqvist et al., 1998). We presented inverse noise faces equally on the left and right sides of the screen in a random order. We used the same pairs of faces for all participants.

Face representation data processing. We 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. We generated both participant-level classification images and group-level classification images (refer back to the Method section 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, we created ingroup ($n = 100$) and outgroup ($n = 100$) classification images (see Figure 7). Second, to examine the differences between Klee and Kandinsky groups, we created Klee ($n = 100$) and Kandinsky ($n = 100$) classification images collapsed across ingroup and outgroup (see Figure 8). Finally, we 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 9.

Phase 2: Assessing impressions of face representations. In Phase 2, we 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 the eight group-level classification images. To assess relative differences between ingroup and outgroup (Group), Klee and Kandinsky (ART), and Group × ART images, we obtained ratings from three different samples of participants. That is, participants only rated
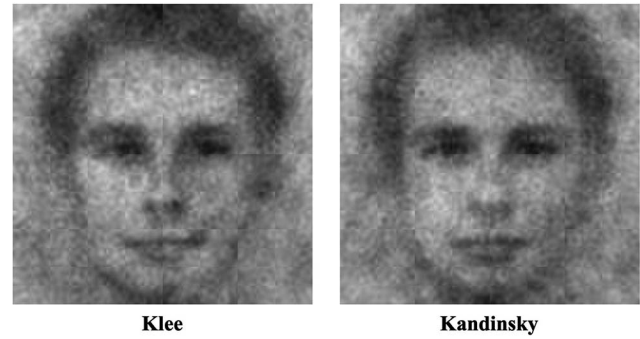
ingroup and outgroup images, Klee and Kandinsky images, or Group × ART images.

Participants. We 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). Fifty 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. The racial and ethnic breakdown of our sample of raters was 103 White, 26 Black, six Latinx, five Asian, one Native American, one Pacific Islander/Hawaiian, and eight multiracial participants. Participants were expected to complete the study in 10 min. All participants did not know about the face categorization stage of the study. They were compensated with $1 for their participation. Participants provided written informed consent approved by the UCSB Human Subjects Committee.

Procedure. After providing informed consent, 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,
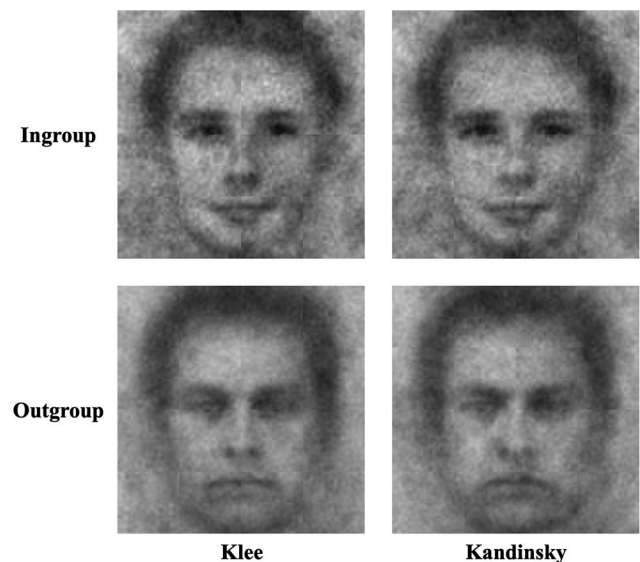

Figure 9. Study 2 Group × ART group-level classification images.

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, we conducted a repeated-measures multivariate analysis of variance (rMANOVA) followed by a univariate analysis of variance for each trait. We 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), $p <$ .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 .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.

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.

Table 4
Study 2 Trait Rating ANOVA Results—Group (ART)—Face Representations

Trait	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

Note. ART = Artistic preference.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 5
Study 2 Trait Rating ANOVA Results—ART—Face Representations

Trait	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

Note. ART = Artistic preference.
+ $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Group × ART. We 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 × ART (Pillai's Trace = .07, F = .75, df = (13, 135), p = .71) were not significant. 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 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.

Discussion

In Study 2a, we investigated the generalizability of our findings from Study 1a to a different minimal group paradigm. First, we replicated the ingroup positivity effect in face representations: Ingroup faces elicited overall more positive trait impressions compared with 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 with the overestimator and underestimator version, the ingroup face elicited even more positive trait impressions than the outgroup face. This is despite the fact that our sample sizes were smaller in this study compared with Study 1a.

We also found some support for the generalizability of the 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. We did

Table 6

Study 2 Trait Rating ANOVA Results—Group × ART—Face Representations

	Klee			Kandinsky			F values		
Trait	Ingroup (SD)	Outgroup (SD)	Cohen's d	Ingroup (SD)	Outgroup (SD)	Cohen's d	Group	ART	Group × 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

Note. ART = Artistic preference.
* $p < .05$. ** $p < .01$. *** $p < .001$.

not expect to find more favorable trait impressions for the Klee group face compared to the Kandinsky group face, so we are hesitant to interpret why this pattern emerged. Nevertheless, our findings still demonstrate that different minimal groups 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 we crossed category labels with group membership, the differences between the Klee group and the Kandinsky group were no longer statistically significant. 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, our findings show that people may infer different traits from category labels in different types of minimal group paradigms, 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).

Study 2b

In Study 2b, we 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 × 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: (a) the minimal group label of the targets (Klee fan or Kandinsky fan); and (b) whether the targets shared the same group membership with them or not (ingroup or outgroup). Because we did not find any effects of minimal group labels on Group × ART face images, we expected to find greater contributions of Group than ART in trait representations of Group × ART images.

Method

Participants. The data from the same 150 participants recruited in Phase 2 of Study 2a were reanalyzed here. Fifty 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 2a for a more detailed description.

Procedure. We followed the same steps of multiple regression RSA outlined in Study 1b to examine contributions of Group and ART in the face categorization task while controlling for each other: (a) we computed pairwise correlations of trait rating data for each group-level classification image; (b) vectorized unique pairwise correlation matrices (i.e., excluding duplicate correlation coefficients); and (c) predicted vectors of Group × ART trait rating data with linear combinations of vectors of corresponding Group and ART trait rating data.

Results

Ingroup Klee. We used ordinary least squares multiple regression to predict the pairwise correlation vector of the trait rating data of the ingroup Klee face image with the linear combination of the correlation vectors of ingroup face trait rating data and Klee face trait rating data. We 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. We 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. We used multiple regression to predict outgroup Klee trait rating data with the linear combination of outgroup face trait rating data and Klee face trait rating data. We 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 ($\beta = .03$, $SE = .07$, $t = .57$, $p = .570$). 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. We used multiple regression to predict ingroup Kandinsky trait rating data with the linear combination of ingroup face trait rating data and Kandinsky face trait rating data. We found that both the ingroup ratings ($\beta = .83$, $SE = .04$, $t = 16.45$, $p < .001$) and Kandinsky ratings ($\beta = .19$, $SE = .08$, $t = 3.74$, $p < .001$) 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. We used multiple regression to predict outgroup Kandinsky trait rating data with the linear combination of outgroup face trait rating data and Kandinsky face trait rating data. We found that the outgroup ratings ($\beta = .88$, $SE = .06$, $t = 15.15$, $p < .001$) were a significant predictor of outgroup Kandinsky ratings, whereas the Kandinsky ratings were not ($\beta = .04$, $SE = .11$, $t = .76$, $p = .45$). 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 Study 2b, we found that for all four Group $\times$ ART face representations, participants seemed to have used the ingroup/outgroup distinction more than the minimal group labels (Klee and Kandinsky) as indicated by larger trait representational similarities between Group $\times$ 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 the representations of the ingroup faces, our current findings are unlike our findings of Study 1b, in that participants in all conditions seemed to have focused more on whether the targets shared their group. Thus, we argue that although minimal group labels can be meaningful when visualizing faces (e.g., the overestimator label in Study 1), minimal group 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 1b apply to the current study. Our interpretations of the 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 the minimal group paradigm are only weakly represented.

Study 2c

Study 2a and 2b showed that there might be some differences between face representations of Klee and Kandinsky groups, but whether the targets were ingroup or outgroup played a bigger role than the minimal group labels, which is in opposition to the findings of Study 1. Furthermore, we found that the magnitude of differences between ingroup and outgroup was larger in this version of the minimal group paradigm, indicating that different degrees of meaningfulness of minimal group labels may moderate intergroup bias in face representations of ingroup and outgroup. In Study 2c, we 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 1c, we expected the algorithms to perform better at 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.

Method

Stimuli. We used 200 participant-level classification images from Phase 1 of Study 2a. Each image had three dimensions: (a) group (ingroup or outgroup); (b) ART (Klee or Kandinsky); and (c) Group $\times$ ART (ingroup-Klee, outgroup-Klee, ingroup-Kandinsky, or outgroup-Kandinsky).

Procedure. We followed the same steps to conduct the machine learning analyses as described in Study 1c. To recap, we (a) vectorized and down sampled pixel intensity data of each image, (b) standardized the data, and (c) performed classification using support vector machines (SVM) with a radial basis function kernel (with default cost and gamma hyperparameters). We used 10-fold cross-validation for each analysis (i.e., classifying between ingroup and outgroup faces, Klee and Kandinsky faces, and Group $\times$ 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. We repeated this step 10 times. We then computed accuracy scores by averaging accuracies from these 10 folds. Next, we tested whether our classifiers performed better than chance by using 1000 permutation tests for each analysis. We 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 $\times$ 2-fold cross-validation paired samples t test (Dietterich, 1998). Please see the Procedure section of Study 1c for more details.

---

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

Results

We were able to classify between ingroup and outgroup images from pixel intensity data significantly better than chance (accuracy = 80.00%, $p < .001$). However, we failed to classify between Klee and Kandinsky group face images better than chance at a .05 significance level (accuracy = 57.00%, $p = .08$). Next, we compared classification accuracies for Group and ART using the 5×2-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 our 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 [19.83, 28.37].

Finally, multiclass SVM results showed that our 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% (one out of four). The confusion matrix of results showed that our 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 Study 2c, we investigated whether the pattern of results found in Study 2a and 2b (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. We were 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., $p = .08$). However, the multiclass results suggest that the aesthetic preference was represented to some extent in the pixel intensity data. Considering both Study 1 and Study 2 together, the fact that differences between face representations of ingroup and outgroup exist across different types of the minimal group paradigm at both individual and group levels suggests that the ingroup positivity effect in face representations is robust 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 1c that the SVM classified between minimal group 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 our finding of more misclassifications within group labels than within ART labels of Group $\times$ ART participant-level classification images. Although descriptive, it seems to suggest 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 3

Studies 1 and 2 served as a deep dive into how the representation of ingroup and outgroup faces in minimal group contexts might be influenced by category labels such as the overestimator/underestimator and the Klee/Kandinsky labels that were previously believed to be arbitrary and trivial. We showed that these labels could matter, and the meaningfulness of a minimal group label may moderate intergroup bias. Although face representations of different groups approximated by the reverse correlation method could reveal how group membership might influence behavior, such as trusting behavior in an economic trust game (Ratner et al., 2014), the implications of the effects of meaningful (or lack thereof) minimal group labels for intergroup behaviors remain unclear. Do people still differentiate between overestimators and underestimators when they do not need to visualize their faces? If so, do they ascribe different traits to these groups? Study 3 sought to address these questions by using the classic Tajfel matrices task and the overestimator/underestimator minimal group paradigm (Tajfel et al., 1971). In this study, we first assigned participants into one of two groups (overestimator or underestimator) and asked them to allocate resources (points) to two anonymous individuals. Critically, the only information participants had was their group membership and that one individual was an overestimator and the other was an underestimator. We then asked our participants to rate overestimator and underestimator on 13 trait dimensions that we used in earlier studies to examine trait impressions elicited by faces of different groups (Oosterhof & Todorov, 2008). By doing so, we examined whether people ascribe different traits to different groups, even when they were not required to visualize their faces. Thus, Study 3 examined (a) whether people discriminate between overestimator and underestimator during a resource allocation task and (b) whether people associate overestimator and underestimator with different traits when they do not need to visually represent their faces.

Method

Participants. We recruited 200 MTurk participants ($M_{\text{age}} = 38.01$, $SD = 10.66$; 103 female, 96 male, one other) via the CloudResearch website (Litman, Robinson, & Abberbock, 2017; www.cloudresearch.com/) to complete an online survey administered through Qualtrics (www.qualtrics.com). Because we used the methods and procedures of Loersch and Arbuckle (2013), we attempted to recruit a comparable number of participants as them ($n = 76$). We simply rounded this number to 100, and because we were interested in effects of both Group and NEST (Loersch & Arbuckle, 2013 only examined the group effect), we then doubled the sample size. We preregistered this sample size before data collection (https://osf.io/s9243/?view_only=92afae84a38548e8a9412e8353f30905). The racial and ethnic breakdown of our sample was 137 White, 25 Asian, 17 Black, 11 Latino/Hispanic, nine multiracial, and one other participant. Participants were expected to complete the study in 15 min but given up to 30 min to finish. All participants were compensated with $2 for their participation. Participants provided written informed consent approved by the UCSB Human Subjects Committee.

Procedure. The current study used the resource allocation task from Study 1 of Loersch and Arbuckle (2013); however, we used a minimal group paradigm to assign participants to different groups instead of using real-world groups. First, as was the case with Study 1a, we used the NEST, a variant of the classic "dot estimation" procedure (Experiment 1 from Tajfel et al., 1971), to assign participants to either the overestimator or underestimator group. See the NEST procedural details from Study 1a for more information about this group manipulation.

Resource allocation task. After completing the NEST, participants completed a series of six Tajfel matrices adapted from Loersch and Arbuckle (2013). Each matrix consists of 13 columns and two rows. On each trial, participants chose a column of two numbers to indicate points they would allocate to two anonymous individuals. Critically, the only information they had about these individuals was their numerical estimation style (overestimator or underestimator). If the target individual shared the same group membership as the participant, then they were allocating points to an ingroup member, if the target had a different group membership, then they were allocating points to an outgroup member. Although different point options allowed us to examine different strategies people could have used in this task, such as parity (allocating equal points) or maximizing group differences (giving more points to the ingroup member than the outgroup member), we simply summed the overall amount participants allocated to each individual and examined the effects of target category label (overestimator or underestimator), whether the target shared the group membership with the participants or not (ingroup or outgroup), and the interaction between the two. By doing so, we examined whether people showed ingroup favoritism (i.e., on average giving more points to the ingroup member than the outgroup member) as well as whether they allocated different points to overestimator and underestimator individuals regardless of their own group membership.

Trait ratings. After completing the resource allocation task, participants rated both overestimator and underestimator on the same 13 trait dimensions that we used to assess trait impressions elicited by faces of different groups in Phase 2 of Study 1a and Study 2a (i.e., trustworthy, attractive, dominant, caring, sociable, confident, emotionally stable, responsible, intelligent, aggressive, mean, weird, and unhappy; Oosterhof & Todorov, 2008). We simply asked them to rate a typical overestimator and a typical underestimator on the 13 traits. Ratings were made on scales from 1 (*not at all*) to 7 (*extremely*). The order of each trait presentation was random.

Likability ratings. Lastly, we asked our participants how they felt about overestimators and underestimators on a 5-point scale from 1 (*very negative*) to 5 (*very positive*).

Results

Resource allocation task. To examine the effects of minimal group labels on how people allocate resources to various individuals, we conducted a mixed-design ANOVA with Group (ingroup, outgroup), NEST (overestimator, underestimator), and the interaction between the two as factors. We found a main effect of Group indicating that participants allocated more points to the ingroup member ($M = 15.72$, $SD = 4.62$) than to the outgroup member ($M = 14.17$, $SD = 4.95$), $F(1, 198) = 46.50$, $p < .001$, Cohen's $d = .42$. We also found a main effect of NEST: Participants

allocated more points to the overestimator ($M = 15.23$, $SD = 4.74$) than to the underestimator ($M = 14.65$, $SD = 4.94$), $F(1, 198) = 6.62$, $p = .01$, Cohen's $d = .08$. The interaction between Group and NEST was not significant, $F(1, 198) = .26$, $p = .61$.

Trait ratings. We used rMANOVA to test the effects of Group, NEST, and the interaction between the two on trait ratings of overestimator and underestimator. Significant multivariate effects were found for Group (Pillai's Trace $= .08$, $F = 2.73$, $df = (13, 383)$, $p < .001$) and NEST (Pillai's Trace $= .25$, $F = 9.92$, $df = (13, 383)$, $p < .001$). The interaction term did not yield a significant multivariate effect (Pillai's Trace $= .05$, $F = 1.63$, $df = (13, 383)$, $p = .08$). We then conducted univariate F tests examining the effects of Group and NEST separately. The results showed that the ingroup was rated more trustworthy, attractive, caring, responsible, and intelligent, whereas the outgroup was rated more mean and unhappy. We also found that the overestimator was rated more dominant, confident, aggressive, and sociable, whereas the underestimator was rated more responsible and unhappy. The univariate F test results including the means, standard deviations, F values, p values, and effect sizes for effects of Group and NEST are presented in Table 7 and 8.

Likability ratings. We conducted a mixed-design ANOVA with Group (ingroup, outgroup), NEST (overestimator, underestimator), and the interaction between the two as factors on people's ratings of overestimator and underestimator. We found a main effect of Group indicating that participants rated their ingroup ($M = 3.48$, $SD = .63$) more favorably than their outgroup ($M = 3.10$, $SD = .59$), $F(1, 198) = 42.59$, $p < .001$, Cohen's $d = .46$. We did not find a significant main effect of NEST, $F(1, 198) = .19$, $p = .66$. The interaction between Group and NEST was also not significant, $F(1, 198) = .48$, $p = .49$.

Discussion

In Study 3, we investigated the effects of minimal group labels on intergroup behavior when people are not required to visualize faces of different groups. First, we replicated the classic ingroup favoritism finding with the resource allocation task (Tajfel et al., 1971). Specifically, participants allocated more points to a member of their ingroup than a member of their outgroup. More interest-

Table 7

Study 3 Trait Rating ANOVA Results—Group

Trait	Ingroup mean (SD)	Outgroup mean (SD)	F value	Cohen's d
Trustworthy	4.91 (1.20)	4.45 (1.11)	16.21***	.32
Attractive	4.31 (1.06)	4.09 (1.02)	4.46*	.16
Dominant	4.00 (1.44)	3.87 (1.47)	.86	.06
Caring	4.72 (1.13)	4.39 (1.15)	8.36**	.24
Confident	4.70 (1.27)	4.54 (1.42)	1.59	.08
Emotionally stable	4.62 (1.21)	4.39 (1.20)	3.47+	.16
Responsible	4.91 (1.23)	4.37 (1.21)	19.14***	.33
Intelligent	4.96 (1.08)	4.40 (1.09)	26.18***	.45
Aggressive	3.37 (1.58)	3.52 (1.58)	.96	.07
Mean	2.52 (1.37)	2.88 (1.39)	7.22**	.28
Weird	3.02 (1.47)	3.16 (1.41)	.89	.11
Unhappy	2.78 (1.39)	3.06 (1.32)	4.18*	.19
Sociable	4.55 (1.05)	4.42 (1.20)	1.33	.09

+ $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 8
Study 3 Trait Rating ANOVA Results—NEST

Trait	Overestimator mean (*SD*)	Underestimator mean (*SD*)	*F* value	Cohen's *d*
Trustworthy	4.59 (1.18)	4.76 (1.16)	2.22	.11
Attractive	4.28 (1.01)	4.12 (1.08)	2.06	.11
Dominant	4.45 (1.40)	3.42 (1.32)	57.86***	.55
Caring	4.49 (1.14)	4.63 (1.16)	1.48	.10
Confident	5.14 (1.11)	4.10 (1.37)	69.73***	.60
Emotionally stable	4.49 (1.16)	4.53 (1.26)	.14	.03
Responsible	4.49 (1.23)	4.79 (1.26)	5.63*	.17
Intelligent	4.67 (1.08)	4.68 (1.16)	.00	.00
Aggressive	3.98 (1.61)	2.91 (1.36)	51.22***	.55
Mean	2.79 (1.40)	2.62 (1.37)	1.45	.12
Weird	3.08 (1.45)	3.10 (1.43)	.01	.01
Unhappy	2.79 (1.28)	3.05 (1.43)	3.88*	.19
Sociable	4.75 (1.03)	4.21 (1.15)	24.29***	.38

Note. NEST = Numerical Estimation Style Test.
$^+ p < .10.$ $^* p < .05.$ $^{**} p < .01.$ $^{***} p < .001.$

ingly, participants also allocated more points to the overestimator person than to the underestimator person, regardless of their own group membership (i.e., no significant interaction between Group and NEST). It is also important to note that allocating more resources did not mean that the overestimator was liked more than the underestimator. That is, although people rated their ingroup as more likable than their outgroup, there was no difference in the likability ratings between overestimator and underestimator.

The resource allocation results suggest that participants might have imbued overestimators and underestimators with different qualities and this influenced their decisions to allocate more points to overestimator than underestimator, as opposed to simply favoring overestimator over underestimator. The trait rating data support this interpretation. First, we found ingroup positivity bias in trait rating data similar to the face representation studies: People rated their ingroup as more trustworthy, attractive, caring, responsible, and intelligent, whereas they rated their outgroup as more mean and unhappy. Second, we found differences between overestimator and underestimator on a number of traits, mirroring the findings of Study 1a: People rated overestimators as more dominant, confident, aggressive, and sociable, and rated underestimators as more unhappy. These findings suggest that the traits that people associate with overestimators are different from the traits that they associate with underestimators, and further supports the idea that people imbue meaning to the overestimator and underestimator labels. It is possible that differential trait attributions to the overestimator and underestimator labels could have led to differential resource allocations. Future research should be designed to examine the possibility of this mediation pathway. Despite an influence of NEST labels on trait impressions and resource allocations, there was not a significant interaction between Group and NEST (i.e., $p = .08$ for trait ratings; $p = .61$ for resources allocation).

Study 4

Study 3 showed that people differentiate between the overestimator and underestimator labels even when they do not need to visually represent their faces. Is this phenomenon specific to the overestimator/underestimator paradigm? Do people differentiate

minimal group labels that have less inductive potential (e.g., Klee/Kandinsky)? The final study of the current research investigated the generalizability of our findings from Study 3 with the Klee/Kandinsky paradigm (Experiment 2 from Tajfel et al., 1971). Study 4 used the same set of methods from Study 3 to examine whether people differentiate between people who like Klee paintings and those who like Kandinsky paintings during a resource allocation task. We again used the Tajfel matrices to assess how people allocate resources to others who are Klee or Kandinsky fans. Because we found less distinction between the Klee and Kandinsky groups in Study 2 compared with the overestimator and underestimator groups in Study 1, we expected to find little to no difference in how people allocate resources to Klee and Kandinsky groups. We also expected to find stronger ingroup favoritism (i.e., more discrimination between ingroup and outgroup) because lack of meaning in minimal group labels was associated with larger differences between mental representations of ingroup and outgroup faces in Study 2.

Method

Participants. We recruited 199 MTurk participants ($M_{age} = 37.03$, $SD = 10.97$; 102 female, 96 male, 1 other) via the CloudResearch website (www.cloudresearch.com) to complete an online survey administered through Qualtrics (www.qualtrics.com). Using the same rationale as Study 3, we tried to recruit 200 participants but due to an unidentifiable error in the recruitment process we ended up with 199 participants. We also preregistered this sample size before data collection (https://osf.io/s9243/?view_only=92afae84a38548e8a9412e8353f30905). The racial and ethnic breakdown of our sample was 137 White, 20 Black, 19 Asian, 13 Latino/Hispanic, nine multiracial, and one other participant. Participants were expected to complete the study in 15 min but were given up to 30 min to finish. All participants were compensated with $2 for their participation. Participants provided written informed consent approved by the UCSB Human Subjects Committee.

Procedure. The current study followed the same procedure as Study 3 except for the version of the minimal group paradigm that assigned participants to different groups. As was the case with Study 2a, we used a classic *aesthetic preference* procedure (Experiment 2 from Tajfel et al., 1971) to assign participants to either the Klee or Kandinsky group.

Resource allocation task. Next, participants allocated points to two anonymous individuals in a series of six Tajfel matrices similar to the ones used in Study 3 except that the only information they had about these individuals was their aesthetic preference (Klee fan or Kandinsky fan). If the target individual shared the same aesthetic preference as the participant, then they were allocating points to an ingroup member, if the target had a different aesthetic preference, then they were allocating points to an outgroup member. We summed the overall amount that participants allocated to each individual and examined the effects of target category label (Klee fan or Kandinsky fan), whether the target shared the same aesthetic preference as the participants or not (ingroup or outgroup), and the interaction between the two. Thus, we examined whether people would show ingroup favoritism as well as whether they allocated different points to Klee and Kandinsky fans regardless of their own aesthetic preference.

Trait ratings. After completing the resource allocation task, participants rated both Klee and Kandinsky fans on the same 13 trait dimensions (i.e., trustworthy, attractive, dominant, caring, sociable, confident, emotionally stable, responsible, intelligent, aggressive, mean, weird, and unhappy; Oosterhof & Todorov, 2008). We simply asked them to rate a typical person who likes paintings of Klee and a typical person who likes paintings of Kandinsky on the 13 traits. Ratings were made on scales from 1 (*not at all*) to 7 (*extremely*). The order of each trait presentation was random.

Likability ratings. Lastly, we asked our participants how they felt about Klee fans and Kandinsky fans on a 5-point scale from 1 (very negative) to 5 (very positive).

Results

Resource allocation task. To examine the effects of minimal group labels on how people allocate resources, we conducted a mixed-design ANOVA with Group (ingroup, outgroup), ART (Klee, Kandinsky), and the interaction between the two as factors. We found a main effect of Group indicating that participants allocated more points to the ingroup member ($M = 16.19$, $SD = 4.69$) than to the outgroup member ($M = 13.69$, $SD = 5.04$), $F(1, 197) = 104.523$, $p < .001$, Cohen's $d = .63$. We did not find a main effect of ART, $F(1, 197) = 1.88$, $p = .17$. The interaction between group and ART was also not significant, $F(1, 197) = .77$, $p = .38$.

Trait ratings. We used rMANOVA to test the effects of Group, ART, and the interaction between the two on trait ratings of overestimator and underestimator. The only significant multivariate effect was found for Group (Pillai's Trace $= .10$, $F = 3.37$, $df = (13, 381)$, $p < .001$). The multivariate effect of ART was not significant (Pillai's Trace $= .01$, $F = .39$, $df = (13, 381)$, $p = .97$). The multivariate effect of the interaction term was also not significant (Pillai's Trace $= .03$, $F = .82$, $df = (13, 381)$, $p = .64$). We thus followed up with univariate F tests examining the effects of only Group. The results showed that the ingroup was rated more trustworthy, attractive, caring, confident, emotionally stable, responsible, and intelligent, whereas the outgroup was rated more aggressive and unhappy. The univariate F test results including the means, standard deviations, F values, p values, and effect sizes for effects of Group are presented in Table 9.

Likability ratings. We conducted a mixed-design ANOVA with Group (ingroup, outgroup), ART (Klee fan, Kandinsky fan), and the interaction between the two as factors on people's ratings of Klee fans and Kandinsky fans. We found a main effect of Group indicating that participants rated their ingroup ($M = 3.94$, $SD = .73$) more favorably than their outgroup ($M = 3.35$, $SD = .69$), $F(1, 197) = 86.52$, $p < .001$, Cohen's $d = .66$. We did not find a significant main effect of ART, $F(1, 197) = .11$, $p = .74$. The interaction between Group and ART was also not significant, $F(1, 197) = .03$, $p = .86$.

Discussion

Study 4 investigated whether people differentiate between the Klee and Kandinsky groups during a resource allocation task as they do with overestimator and underestimator (Study 3) even when people are not required to represent faces of different groups. First, we replicated the classic ingroup favoritism finding with the resource allocation task again

Table 9

Study 4 Trait Rating ANOVA Results—Group

Trait	Ingroup mean (SD)	Outgroup mean (SD)	F value	Cohen's d
Trustworthy	5.07 (1.13)	4.54 (1.14)	21.27***	.32
Attractive	4.72 (1.08)	4.22 (1.04)	22.39***	.34
Dominant	3.78 (1.43)	3.66 (1.37)	.80	.06
Caring	5.00 (1.12)	4.57 (1.14)	14.36***	.27
Confident	4.94 (1.15)	4.70 (1.15)	4.43*	.14
Emotionally stable	4.88 (1.22)	4.53 (1.17)	8.90**	.21
Responsible	5.06 (1.14)	4.57 (1.15)	17.94***	.30
Intelligent	5.21 (1.12)	4.75 (1.23)	15.61***	.26
Aggressive	2.80 (1.40)	3.15 (1.48)	5.79*	.18
Mean	2.40 (1.37)	2.63 (1.40)	2.82+	.12
Weird	3.20 (1.57)	3.50 (1.58)	3.79+	.13
Unhappy	2.60 (1.37)	2.91 (1.43)	5.05*	.15
Sociable	4.75 (1.21)	4.57 (1.20)	2.14	.10

$^+ p < .10.$ $^* p < .05.$ $^{**} p < .01.$ $^{***} p < .001.$

(Tajfel et al., 1971): People allocated more points to the individual who shared the same artistic preference as them (ingroup) than to the individual who did not share the same artistic preference (outgroup). However, participants did not discriminate between the Klee and Kandinsky fans when allocating resources. As predicted, however, this was related to stronger ingroup favoritism on the resource allocation task in Study 4 as indicated by the larger effect size ($d = .63$) compared with the results on this same task in Study 3 when participants were influenced by the minimal group labels ($d = .42$). We also replicated the finding from Study 3 that people liked their ingroup more than their outgroup. The effect size from this finding was again greater ($d = .66$) than the one found in Study 3 ($d = .46$), providing another piece of evidence that when people do not differentiate between minimal group labels, they may differentiate between ingroup and outgroup more, thus showing greater intergroup bias.

Furthermore, we found no effect of minimal group labels on any of the 13 traits, indicating that participants did not differentiate between Klee and Kandinsky fans when they were not representing their faces. However, we replicated ingroup positivity effects in these trait rating data. That is, participants rated their ingroup as more trustworthy, attractive, caring, confident, emotionally stable, responsible, and intelligent, whereas rated their outgroup as more aggressive and unhappy. We not only found significant differences between ingroup and outgroup on a greater number of traits in this paradigm compared with the overestimator/underestimator paradigm, but also that a majority of the significant traits in both paradigms yielded larger effect sizes in the current version of the minimal group paradigm than the one used in Study 3. These findings further support the idea that when the meaning of the minimal group labels is diminished, people may focus more on the ingroup/outgroup distinction, which leads to greater intergroup bias.

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 article was to empirically examine whether the classic minimal group category labels that have been used in many research studies over the past 50 years are ever imbued with meaning and whether such associations have

consequences for various intergroup responses. We probed this question across four studies.

We first used an overestimator/underestimator 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. Beyond successfully replicating their ingroup positivity main effect, we showed that category labels also mattered, in that, faces of overestimators and underestimators are represented differently. Our 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. We then conducted representational similarity analysis on the trait rating data, and found that the overestimator label contributed to the pattern of trait ratings more so than the ingroup/outgroup distinction, but this was not the case for the underestimator label. We 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 our algorithm's better accuracy for classifying the NEST labels than the ingroup/outgroup labels.

Next, we sought to understand the generalizability of the findings from Study 1. We did so 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. We 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 far less than the ingroup/outgroup distinction. This in turn was accompanied by larger differences between ingroup and outgroup. We corroborated these findings with the machine learning analysis and showed that there were larger differences between faces of ingroup and outgroup than Klee and Kandinsky groups at the participant level.

Despite using multiple analysis techniques to demonstrate effects of minimal group labels on category representation with different minimal group paradigms, we next sought to examine whether these category label effects translated into meaningful overt evaluations, impressions, and behaviors. To this end, we replicated Tajfel et al.'s (1971) classic overestimator/underestimator minimal group study with a resource allocation task adapted from Loersch and Arbuckle (2013). Consistent with Tajfel et al.'s (1971) original finding, participants allocated more resources to their minimal ingroup than outgroup. However, we also found that people allocated more resources to an overestimator than to an underestimator, although they did not evaluate overestimators more favorably than underestimators. We also showed that people associated different traits with overestimators and underestimators, suggesting that people imbue minimal group labels with meaning even when they do not need to visually represent faces.

In our final study, we examined the effects of the Klee and Kandinsky labels on resource allocations, given that logically the Klee and Kandinsky labels provide less to read into than do overestimator and underestimator labels. We found that participants did not differentiate between Klee fans and Kandinsky fans when it came to allocating resources, evaluating them, and ascribing different traits. This lack of differentiation between the Klee and Kandinsky groups was accompanied by stronger ingroup favoritism on the resource allocation task as well as more favorable evaluations of ingroup than outgroup. It is intriguing that the reverse correlation procedure used in Study 2 identified representational differences between the Klee and Kandinsky groups, and these differences did not emerge in explicit trait ratings, evaluations, and behavioral responses, but studies consistently revealed stronger intergroup bias in the Klee/Kandinsky version than the overestimator/underestimator version of the minimal group paradigm.

What does this all mean for Tajfel et al.'s (1971) foundational assumption? It appears that category labels are flimsy in some ways, but not in others. On one hand, intergroup bias reliably emerges irrespective of category label. On the other hand, category labels are represented differently and can moderate intergroup bias in representations when labels are high in inductive potential, such as the overestimator/underestimator distinction. Moreover, when high in inductive potential, these labels can carry more weight in the overall representations than the ingroup/outgroup distinction. However, the impact on downstream impressions, evaluations, and behavior appears to be limited. Impressions of and resource allocations to overestimators and underestimators differed, but these label effects did not significantly interact with the group effects, which supports the validity of the intergroup bias interpretations in the literature.

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 old adage "if it ain't broke, don't fix it" probably contributed to this inertia, given that the paradigm has been successful, and to our knowledge, minimal group research has not suffered from replicability problems that have plagued other psychological paradigms. From a methodological standpoint, however, our work 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.

Our research also provides insight into how the implied meaningfulness of the labels influences various forms of intergroup bias. A priori, it was not obvious to us 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 paradigm should lead to more intergroup bias than the Klee/Kandinsky version because the inductive potential of the dot estimation labels provides a rationale to identify with one's ingroup. On the other hand, to the extent that one is reading into the dot estimation labels (particularly the overesti-

mator label) then these associations could obscure the ingroup/ outgroup distinction. Our 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, as we expand on below, many real-world novel groups actually have labels that imply characteristics. To the extent that a researcher is interested in modeling this dynamic in the laboratory, 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. We found intergroup effects on all of the dependent variables we examined. Moreover, on the resource allocations and explicit ratings we did not see significant interaction effects between label and group. This lack of an interaction suggests that label effects are less of a concern when outcome variables of interest are behavioral and self-reported impressions versus outcomes that are more sensitive to mental representations of category information (e.g., face representations using the reverse correlation technique).

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 we told participants that numerical estimation style was not related to any other cognitive tendencies or personality traits. Klee and Kandinsky were also represented differently even though we gave participants the same instructions, albeit this distinction was represented to a lesser extent than ingroup and outgroup. Why would people ignore the instructions designed to constrain their inferences about the underlying essence of the groups? As we state earlier, Grice's (1975) Maxim of Quality states that people generally believe what they are told. From this vantage, telling participants that they should not read into the labels should be sufficient. However, 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, Schmidt, Bernhardt, Dierksmeier, & Banse, 2011; Payne, Cheng, Govorun, & Stewart, 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 peculiar, though, 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, agreeable-

ness, and conscientiousness, whereas names with voiceless stop phonemes are associated with high extraversion (Sidhu, Deschamps, Bourdage, & Pexman, 2019). Maybe the way that category labels are pronounced can bias trait inferences. Additional research is necessary to explore these possibilities.

Our research only focused on the two most famous minimal group paradigms. It was beyond the scope of our research to catalog effects of all possible minimal group labels. Although future studies are necessary to understand how broadly the current findings generalize, our evidence of category label effects suggests that a revision is necessary to the account of 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 ingroup versus outgroup classification but 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. We think it is likely 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 to researchers if their participants attempt to read into the category labels and infer meaning from them.

Significance for Reverse Correlation Research in Social Psychology

The goal of the current research was to understand whether category labels influence minimal group responses. Because the reverse correlation method is sensitive to category representation it seemed like an ideal tool to use in our research. In the process of maximizing the utility of this method to answer our research questions, we developed several novel ways of analyzing reverse correlation data. To our knowledge, our work is the first to use 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 we used could prove useful to other social psychologists addressing different research questions with reverse correlation methods. We 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). We are 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.

Why It Matters if Category Labels Matter

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.

If 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 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. Moreover, the type of processing one is engaged in while thinking about such groups could influence the extent to which category labels are functional for serving people's inferential goals.

One of the clearest real-world implications is for visualizing novel group members because the reverse correlation studies revealed the most pronounced category label effects. An example of such a situation is corresponding with another person electronically without knowing what they look like. 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. It is also frequently the case that we are tasked with finding someone when we do not know what they look like. In these situations, we might know something about their group membership and use this knowledge as a cue to infer what they look like. Our work suggests that in both of these 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. 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.

How category labels influence intergroup behavior and impressions during everyday life when visualization is not required is less clear. We did not find evidence that intergroup bias in such circumstances are moderated by category labels. However, the main effect of category label on the resource allocation task when participants were divided into overestimators and underestimators suggests that category labels can have important effects on behavior. For instance, when separated into novel groups in real-world situations, distinctions that signify group membership could imply traits about dominance or likability in ways that affect behavior distinct from mere group membership.

Whether subtle cues like category labels and other attributes 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 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 find 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. As we state earlier, most groups are not labeled by random number or letter strings. Their names contribute to their identity. Thus, category labels could consciously or nonconsciously provide associations that feed intergroup differentiation.

Conclusion

The minimal group paradigm gained prominence because it showed that people discriminate even when group boundaries are meaningless. Our 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. This could especially be the case when people are tasked with information processing goals that are most easily accomplished with access to concrete information (e.g., visualization). By challenging conventional understanding of the minimal group paradigm, our work

provides new insight into the circumstances when category labels might influence intergroup responses.

References

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,* 384–399. http://dx.doi.org/10.1037/0022-3514.59.3.384

Asch, S. E., & Zukier, H. (1984). Thinking about persons. *Journal of Personality and Social Psychology, 46,* 1230–1240. http://dx.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,* 789–799. http://dx.doi.org/10.1037/0022-3514.81.5.789

Augoustinos, M., & Rosewarne, D. L. (2001). Stereotype knowledge and prejudice in children. *British Journal of Developmental Psychology, 19,* 143–156. http://dx.doi.org/10.1348/026151001165912

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,* 706–712. http://dx.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,* 27–52. http://dx.doi.org/10.1002/ejsp.2420030103

Blank, H. (1997). Cooperative participants discriminate (not always): A logic of conversation approach to the minimal group paradigm. *Current Research in Social Psychology, 2,* 38–49.

Blank, H. (2004). Conversation logic effects in the minimal group paradigm: Existent but weak. *Current Research in Social Psychology, 10,* 84–103.

Brewer, M. B., & Silver, M. (1978). Ingroup bias as a function of task characteristics. *European Journal of Social Psychology, 8,* 393–400. http://dx.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,* 333–361. http://dx.doi.org/10.1080/10463283.2017.1381469

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review, 64,* 123–152. http://dx.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,* 3–5. http://dx.doi.org/10.1177/1745691610393980

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

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. http://dx.doi.org/10.1016/j.jesp.2014.06.007

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,* 523–543. http://dx.doi.org/10.1521/soco.2016.34.6.523

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

Cosmides, L., Tooby, J., & Kurzban, R. (2003). Perceptions of race. *Trends in Cognitive Sciences, 7,* 173–179. http://dx.doi.org/10.1016/S1364-6613(03)00057-3

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56,* 5–18. http://dx.doi.org/10.1037/0022-3514.56.1.5

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10,* 1895–1923. http://dx.doi.org/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 & Personality Science, 3,* 562–571. http://dx.doi.org/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,* 978–980. http://dx.doi.org/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,* 999–1014. http://dx.doi.org/10.1037/a0023026

Dunham, Y., Baron, A. S., & Carey, S. (2011). Consequences of "minimal" group affiliations in children. *Child Development, 82,* 793–811. http://dx.doi.org/10.1111/j.1467-8624.2011.01577.x

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

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,* 188–205. http://dx.doi.org/10.1037/0022-3514.80.2.188

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.), *Syntax and semantics, Vol. 3, Speech Acts* (pp. 41–58). New York, NY: Academic Press.

Hastie, R. (1984). Causes and effects of causal attribution. *Journal of Personality and Social Psychology, 46,* 44–56. http://dx.doi.org/10.1037/0022-3514.46.1.44

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

Hugenberg, K., & Corneille, O. (2009). Holistic Processing Is Tuned for In-Group Faces. *Cognitive Science, 33,* 1173–1181. http://dx.doi.org/10.1111/j.1551-6709.2009.01048.x

Imhoff, R., Schmidt, A. F., Bernhardt, J., Dierksmeier, A., & Banse, R. (2011). An inkblot for sexual preference: A semantic variant of the Affect Misattribution Procedure. *Cognition and Emotion, 25,* 676–690. http://dx.doi.org/10.1080/02699931.2010.508260

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

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2,* 1–28. http://dx.doi.org/10.3389/neuro.06.004.2008

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "seizing" and "freezing." *Psychological Review, 103,* 263–283. http://dx.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,* 522–544. http://dx.doi.org/10.1037/0033-2909.129.4.522

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,* 1465–1467. http://dx.doi.org/10.1126/science.aap8731

Lindström, B., Selbing, I., Molapour, T., & Olsson, A. (2014). Racial bias shapes social reinforcement learning. *Psychological Science, 25,* 711–719. http://dx.doi.org/10.1177/0956797613514093

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49,* 433–442.

Loersch, C., & Arbuckle, N. L. (2013). Unraveling the mystery of music: Music as an evolved group process. *Journal of Personality and Social Psychology, 105,* 777–798. http://dx.doi.org/10.1037/a0033691

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

Maddox, K. B. (2004). Perspectives on racial phenotypicality bias. *Personality and Social Psychology Review, 8,* 383–401. http://dx.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,* 209–226. http://dx.doi.org/10.1207/s15516709cog2802_4

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44,* 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

Miller, S. L., Maner, J. K., & Becker, D. V. (2010). Self-protective biases in group categorization: Threat cues shape the psychological boundary between "us" and "them". *Journal of Personality and Social Psychology, 99,* 62–77. http://dx.doi.org/10.1037/a0018086

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316. http://dx.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,* 231–259. http://dx.doi.org/10.1037/0033-295X.84.3.231

Ojala, M., & Garriga, G. C. (2010). Permutation Tests for Studying Classifier Performance. *Journal of Machine Learning Research, 11,* 1833–1863. http://dx.doi.org/10.1109/ICDM.2009.108

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 11087–11092. http://dx.doi.org/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,* 77–89. http://dx.doi.org/10.1006/jesp.1999.1399

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,* 277–293. http://dx.doi.org/10.1037/0022-3514.89.3.277

Pinter, B., & Greenwald, A. G. (2010). A comparison of minimal group induction procedures. *Group Processes & Intergroup Relations, 14,* 81–98. http://dx.doi.org/10.1177/1368430210375251

Przybylinski, E., & Andersen, S. M. (2013). Short-circuiting transference of past relationships using implementation intentions. *Journal of Experimental Social Psychology, 49,* 566–572. http://dx.doi.org/10.1016/j.jesp.2012.09.003

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,* 914–930. http://dx.doi.org/10.1177/0022022117702975

Rabbie, J. M., & Horwitz, M. (1969). Arousal of ingroup-outgroup bias by a chance win or loss. *Journal of Personality and Social Psychology, 13,* 269–277. http://dx.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,* 298–301. http://dx.doi.org/10.1016/j.jesp.2012.10.017

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,* 897–911. http://dx.doi.org/10.1037/a0036498

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

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382–439. http://dx.doi.org/10.1016/0010-0285(76)90013-X

Rothbart, M., & Taylor, M. (1992). Category labels and social reality: Do we view social categories as natural kinds? In G. R. Semin & K. Fiedler (Eds.), *Language, interaction and social cognition* (pp. 11–36). Thousand Oaks, CA: Sage.

Scholkopf, B., Kah-Kay, Sung., Burges, C. J. 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,* 2758–2765. http://dx.doi.org/10.1109/78.650102

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,* 1595–1614. http://dx.doi.org/10.1037/xge0000662

Stolier, R. M., & Freeman, J. B. (2016). Neural pattern similarity reveals the inherent intersection of social categories. *Nature Neuroscience, 19,* 795–797. http://dx.doi.org/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,* 197–200. http://dx.doi.org/10.1016/j.tics.2017.12.003

Tajfel, H. (1970). Experiments in Intergroup Discrimination. *Scientific American, 223,* 96–103.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology, 1,* 149–178. http://dx.doi.org/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.

Van Bavel, J. J., & Cunningham, W. A. (2009). Self-categorization with a novel mixed-race group moderates automatic social and racial biases. *Personality and Social Psychology Bulletin, 35,* 321–335. http://dx.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,* 1131–1139. http://dx.doi.org/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,* 3343–3354. http://dx.doi.org/10.1162/jocn_a_00016

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

Weisbuch, M., Pauker, K., Adams, R. B., Jr., Lamer, S. A., & Ambady, N. (2017). Race, power, and reflexive gaze following. *Social Cognition, 35,* 619–638. http://dx.doi.org/10.1521/soco.2017.35.6.619