Running Head: GROUP RESEMBLANCE

The Faces of Group Members Share Physical Resemblance

Eric Hehman[1], Jessica K. Flake[2], Jonathan B. Freeman[3]

[1]Department of Psychology, Ryerson University, Toronto, ON, M5B 2K3
[2]Department of Psychology, York University, Toronto, ON, M3J 1P3
[3]Department of Psychology, New York University, New York, NY, 10017

Corresponding Author:
    Eric Hehman, Department of Psychology, Ryerson University, E-mail: eric.hehman@psych.ryerson.ca

World count: 7810

Abstract

Perceivers form strong inferences of disposition from others' facial appearance, and these inferences guide a wide variety of important behaviors. The current research examines the possibility that similar-looking individuals are more likely to form groups with one another. We do so by testing a necessary downstream consequence of this process, examining whether the faces of individuals within groups more physically resemble one another than those in other groups. Across six studies we demonstrate that individuals' group membership can be accurately classified both from ratings of members' faces, and from direct measurement of members' faces. Results provide insight into how affiliative groups initially form and maintain membership over time, as well as the perception of homogeneity of groups.

Keywords: person perception, impression formation, face perception, group formation, intragroup dynamics

The Faces of Group Members Share Physical Resemblance

People like similar others. This concept, known as homophily, is a long-established psychological phenomenon and is a core process by which individuals form groups with others. The extent to which facial appearance plays a role in this group formation process, however, has remained unexamined, and it may be a critical mechanism by which individuals form groups. The current research posits that individuals form groups partly based on inferences from facial appearance, and tests this hypothesis by examining a necessary downstream consequence of this possibility: that individuals within groups physically resemble one another.

Individuals prefer to associate with others perceived to have similar interests and beliefs (for review, see McPherson, Smith-Lovin, & Cook, 2001). Friendships, romantic relationships, and groups form around these similarities, as individuals are happier around others with perceived similar interests and opinions (Caspi & Herbener, 1990; Mackinnon, Jordan, & Wilson, 2011). Critical to the present work, it is the *perception* of similarity that is crucial in determining liking (Lee & Bond, 1998). Because perceived similarity leads to group formation, facial appearance may play an important and yet relatively overlooked role in determining group membership.

Facial appearance is a critical initial factor influencing how people are perceived. Models of person perception hold that from a target's facial features perceivers form strong inferences about the target (Brunswik, 1952; Kenny, 1991; Oosterhof & Todorov, 2008; Zebrowitz, Fellous, Mignault, & Andreoletti, 2003), such as inferences about targets' opinions and dispositions (i.e., friendly). While these inferences are not always accurate (Jenkins, White, Van Montfort, & Burton, 2011; Rule, Krendl, Ivcevic, & Ambady, 2013; Todorov & Porter, 2014), it is impressive how often perceivers largely agree (Berry, 1991; Kenny & Albright, 1987; Moskowitz, 1990).

Importantly, these perceived characteristics exert a powerful influence on perceptions of, and behavior towards, target individuals (Brewer, 1988; Dovidio, Kawakami, Johnson, Johnson, & Howard, 1997; Neuberg & Fiske, 1987). Indeed, research has well documented how facial appearance influences outcomes of considerable societal consequence, even when a large amount of ostensibly more valid, objective information is present, such as in judicial decisions (Wilson & Rule, 2015; Zebrowitz & McDonald, 1991) or electoral outcomes (Hehman, Carpinella, Johnson, Leitner, & Freeman, 2014; Todorov, Mandisodza, Goren, & Hall, 2005).

We hypothesize that facial appearance additionally contributes to determining group membership. In the present research, we focus on elective groups in which individuals can seek membership and be accepted (or rejected) by the group. This is in constrast to groups with heritable components (e.g., gender, race), or groups defined by physical characteristics (e.g., physically disabled). An extensive body of research has explored the many ways individuals differentially process pre-existing group boundaries, such as in-group and out-group distinctions, and group memberships such as race, gender, or age (Allport, 1954; Bernstein, Young, & Hugenberg, 2007; Cikara, Botvinick, & Fiske, 2011; Dovidio, Kawakami, & Gaertner, 2002; S. L. Gaertner & Dovidio, 2000; Gray et al., 2014; Hehman, Mania, & Gaertner, 2010; Kubota, Banaji, & Phelps, 2012; Ratner & Amodio, 2013; Stolier & Freeman, 2016). Other research has explored minimal group paradigms or group distinctions such as political or athletic affiliation (Deegan, Hehman, Gaertner, & Dovidio, 2015; Hewstone & Swart, 2011; Hornsey & Hogg, 2000; Mummendey & Wenzel, 1999; Van Bavel & Cunningham, 2009), yet how these more elective and affiliative groups form and are maintained in the first place is less understood. Here, we tested whether individuals within groups physically more resemble one another than individuals in other groups.

Facial appearance may play a crucial role in elective group membership for two reasons. First, individuals may seek out groups perceived as similar. In this case, the perception of similarity may come from inferences of disposition based on the facial appearance of existing group members, and because affiliating with similar others is more enjoyable, seek out membership in these groups (McPherson et al., 2001). There is some limited evidence for this possibility, as people are more likely to approach, interact with, and form romantic relationships with others similar in attractiveness (Alvarez & Jaffe, 2004; Halberstadt et al., 2016), trust similar-looking others (DeBruine, 2002), and sit closer to others with similar length hair and hair color (Mackinnon et al., 2011).

Further, groups may be more likely to accept candidates perceived as similar. When groups are evaluating a potential candidate for acceptance into the group, they are typically making decisions with limited information. In this context, appearance is particularly influential (Ambady & Rosenthal, 1992; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2014). Insofar as individuals within a group use facial appearance as a proxy for underlying personality, groups may make acceptance decisions based on the appearance of potential candidates. Indeed, individuals with faces appearing more physically strong are more likely to be selected for group membership in physically competitive contexts (Hehman, Leitner, Deegan, & Gaertner, 2015). In affiliative groups, individuals who physically resemble existing group members might be more likely to be accepted into the group because group members will be inclined toward inferring they are similar in disposition, being more likely to share the values of the group and get along best with existing group members.

Thus, these two mechanisms would operate under the same priniciple of homophily in determining that similar-looking individuals are more likely to form groups with one another,

though the locus of this may lie at both the candidate- or the group-level. These mechanisms likely act simultaneously and in concert, and parsing the two is beyond the scope of the current research. Because both would lead to the same outcome, our goal was to test the initial foundation for this model, that the faces of group members resemble one another. We hypothesized that should facial appearance predict group membership, either through candidates seeking membership in a similar-looking group, or groups accepting similar-looking candidates, then the individuals within these elective groups should more physically resemble one another than individuals in other groups.

We test this possibility by considering whether it is possible to accurately classify any individual's specific group membership from measurements of their face alone. In the present work, these measurements take two forms: perceiver ratings of social impressions and measurements from three-dimensional models. First, individuals who physically resemble one another should elicit similar social perceptions (e.g., friendly, competent), though this method of testing physical resemblance is indirect. Alternatively, a more direct approach would be to take numerous measurements of a face. Faces that physically resemble one another should be similar in measurement. If group memberships can be accurately classified from these facial measurements, it would provide strong evidence that individuals within groups physically resemble one another more so than those in other groups.

Across six studies we take both approaches described above and find support for our hypotheses. First, in the Supplementary Materials we provide a proof of concept simulation, illustrating the hypothesized mechanism (though others are possible, see General Discussion) that a group accepting similar-looking candidates leads to increased overall homogeneity in the facial appearance of group members over time. Then, Studies 1-5 demonstrate that accurate

classification of group membership arises from both social perceptions and morphological features, such that individuals within affiliative groups share more similarity with one another than with the individuals in other groups. Raw data for Studies 1-5B are available for download at [https://osf.io/8tjg5/], and code for analysis is available in the Supplementary Materials.

Study 1

Study 1 was an initial test of our primary hypothesis, examining whether individuals in pre-existing friendship groups might physically resemble one another such that their appearance might elicit similar enough social perceptions that their membership might be accurately classified.

Methods

Stimuli. Twenty-seven participants (9 female, Age M=31.0, SD=8.75) recruited from Mechanical Turk for monetary compensation submitted 4-6 photographs of their same-gender, close friends' public Facebook profile photographs. Because photographs were the unit of analysis, sample size was determined by aiming to collect photographs from enough participants such that at least 100 would be included in the final analysis. Participants were provided examples of suitable and unsuitable photographs. We required that the photograph was of reasonable resolution, the face was not obscured, and that the friends appeared alone. 137 photographs were provided, and we eliminated 12 that did not meet our criteria. The resulting photographs (n=125) were cropped to the face.

Raters and Procedure. 316 individuals from Amazon Mechanical Turk rated these photographs in exchange for payment on factors previous research has demonstrated critical to person perception: attractiveness, intelligence, and physical strength (Hehman, Leitner, & Freeman, 2014; Oosterhof & Todorov, 2008; Sutherland, Young, Mootz, & Oldmeadow, 2014).

In addition, because Facebook users vary in age, we additionally collected ratings of youthfulness. Targets were presented in random order, and participants rated targets along a 1-"Not at all" to 7-"Very" scale. Participants rated all targets on only a single trait, and each photograph was rated by a minimum of 25 participants on each trait. Participants (n=42) with repeated responses or responding in less than 400ms were removed, leaving 247 for analysis (104 female, 10 unreported, Age M=35.4, SD=11.49). Ratings of each photograph were averaged across participants, and the target photograph operated as the unit of analysis.

Analytical Approach. All four averaged ratings of each photograph were submitted to discriminant function analyses. Discriminate function analyses predict a categorical variable (e.g., group membership) by creating linear combinations of multiple facial ratings (i.e., discriminant functions) that best differentiate membership (Field, 2009). These analyses derived discriminant functions to classify to what friendship group a particular target was most likely to belong. Because this is a data-driven approach that optimally explains variance in the sample on which it is trained, one limitation can be "overfitting", or that the solution derived is so specific to idiosyncrasies of the training sample that it may not generalize to other samples (Babyak, 2004). Accordingly, a critical step is cross-validation, or testing the model derived from a training sample (i.e., the training set) on another sample not used to create the model (i.e., the test set) (Efron & Gong, 1983). In the current analyses, we tested the generalizability of all solutions using two separate cross-validation techniques: hold-out and leave-one-out cross-validation.

Hold-out cross-validation randomly splits the data into two pieces, a training set and a test set. The model derived from the training set is then imposed on the test set, and accuracy is assessed. This approach has the advantages of large training and test sets, and thus better

estimates of the model and its accuracy. Because it is run only once, however, it has the disadvantage that accuracy or error may be a spurious result of whichever cases are randomly selected for the training and test sets. With the leave-one-out approach, sometimes called jackknife, the model is repeatedly (such that repetitions $= n$ observations) trained on all observations excluding one, and a classification is repeatedly made for that single excluded case. The average error across all repetitions is computed and used to evaluate the model. This approach has less bias towards overestimating the true expected error than hold-out, but greater variance. Should both approaches together indicate an adequate fit of the model, it is reasonable evidence that the model generalizes beyond the training sets to other similar groups. Cross-validation was conducted in SPSS and in R using the MASS package (Venables & Ripley, 2002).

Successful classification using these derived models during cross-validation is evidence that discriminate functions were successful in determining group membership of individuals in separate groups (or in the current context, that individuals within a group look more similar than those in other groups). Error rate is typically compared to the rate that is expected by chance in a descriptive manner to assess adequacy of fit. However, to statistically determine that the model derived from hold-out cross-validation was adequate and better than chance, we used two different tests: the Maximum Chance Criterion (MCC) and Press' Q statistic (Q). With the MCC approach the proportion of the largest group in the test sample is calculated. A small buffer to chance is then added, such that if the percentage of targets accurately classified is 1.25x greater than this proportion the validity of the discriminant functions is considered satisfactory (Morrison, 1969). With the Q approach, a statistic is calculated based on the number of targets in the test sample, number of targets accurately classified, and total number of groups[1]. Because

this Press' Q statistic is equivalent to a two-tailed chi-square with $df=1$, critical and p-values can be estimated accordingly. If Q is significant then the null, that the classification matrix is equivalent to chance, is rejected. We report both tests throughout.

Results

We have hypothesized that an individual's group membership can be accurately classified from ratings of their appearance alone. Cross-validation with the hold-out approach supported this hypothesis. 42.2% of the individuals in the training set were accurately classified into their friendship groups. Applying this model to the test set, 14.8% of the individuals in the test set were accurately categorized into their friendship groups. Because 27 different friendship groups were involved, expected accuracy due to chance would be approximately 3.7% accuracy[2]. To determine that 14.8% was significantly better than chance, both MCC and Q approaches were examined. Because accuracy was greater than the MCC critical value (accuracy .148 > MCC critical .098), and because the Q statistic was significant ($Q = 22.09, p < .0001$), both tests of the hold-out approach indicated that the model built on the training sample satisfactorily classified the test sample, and thus that this model was generalizable beyond the training sample. Leave-one-out results converged with this conclusion, with averaged accuracy at 15.2%. Thus, both cross-validation techniques indicated that individuals within friendship groups elicit more similar social perceptions than individuals in other friendship groups, and from these social perceptions their group membership can be classified better than chance.

Four discriminant functions derived from the participant ratings were responsible for this accurate classification (Table 1). Discriminant functions are blends of variables that optimally classify group membership, and examination of the functions in the table reveals their composition. For instance, Function 1 explained 42.6% of the variance in social perceptions, and

was a mix of positive loadings of strength and youthfulness, and negative loading of attractiveness. Intelligence did not contribute very much to this function. Function 2, in contrast, explained 26.4% of the variance, and was primarily comprised of attractiveness. These functions are entirely data-driven, and so while examination of the functions can be informative, their composition can be expected to vary across different contexts and types of groups. What is most important to our current hypotheses is not their content, but rather their ability to accurately classify targets to existing groups.

Traits	Discriminant Functions			
	1	2	3	4
Attractive	-1.057	0.915	0.484	-0.098
Intelligent	0.103	-0.305	-0.053	1.037
Strong	1.025	-0.481	0.719	0.041
Youthful	1.427	0.221	-0.214	-0.019
% Variance	42.6	26.4	19.3	11.7

Table 1. Composition of the four functions created by the discriminant function analysis using the hold-out approach in Study 1.

These friendship groups included both male and female targets, and to the extent that target-gender influenced ratings along different traits, gender may have aided in classification accuracy. To ensure gender alone could not account for accurate classification we collected new data and repeated discriminant function analyses within each gender. Results were fully consistent with those reported here. Please see the Supplementary Materials for a full description of these results.

Study 2

While Study 1 provided evidence for our hypothesis that individuals within social groups more resemble one another than individuals in other groups, our conclusions were limited by

some potential artifacts of the photographs used. Mainly, individuals within friendship groups may systematically vary in the quality of photograph, meaning they may vary in low-level attributes that might influence perceivers when forming social perceptions. Study 2 addressed this limitation.

Methods

Stimuli. To eliminate the possibility that low-level stimulus attributes such as luminance, contrast, and spatial frequency were varying systematically across friendship groups (potentially from different quality cameras) and spuriously producing our results, all stimuli were equated on these factors using the SHINE toolbox in MATLAB (Willenbockel et al., 2010).

Raters and Procedure. Participants were again recruited from Mechanical Turk and rated the photographs in a manner identical to Study 1. Participants (n=60) were removed with the same criteria as Study 1, leaving 248 for analysis (99 female, 18 unreported, Age M=34.4, SD=10.76). Ratings were again averaged, with the photograph serving as the unit of analysis. All ratings were submitted to the discriminant function analysis.

Results

Replicating Study 1, cross-validation with the hold-out approach supported our hypotheses. 34.8% of the individuals in the training set were accurately classified into their friendship groups. Applying this model to the test set, 10.1% of the individuals were accurately categorized into their friendship groups. Again, because 27 different friendship groups were involved, expected accuracy due to chance would be approximately 3.7%. The MCC and Q approaches tested whether 10.1% was significantly greater than the 3.7% expected by chance. Both approaches (accuracy .101 > MCC critical .091; Q = 9.47, p =.0021) indicated that the model successfully generalized beyond the training sample. Accuracy with leave-one-out cross-

validation was 16.8%, also well above chance. Thus, replicating Study 1 with stimuli controlling for luminance, contrast, and spatial frequency, from these results we can conclude that individuals within friendship groups elicit more similar social perceptions than individuals in other friendship groups, and from these social perceptions their group membership can be classified. See Supplementary Table 14 for discriminant function composition.

As in Study 1, these friendship groups included both male and female targets. We therefore collected additional data and repeated our analyses within gender to ensure gender alone could not account for accurate classification. Results were fully consistent with those reported here. Please see the Supplementary Materials for a full description of these results.

Study 3

A limitation of Studies 1 and 2 is that we did not place any restrictions on which participants could provide photographs of their friend group members. Therefore, one potential reason classification accuracy was so high may be a high degree of variation in basic perceptual or demographic characteristics. For instance, though we have demonstrated accurate classificaiton is possible while controlling for gender, other factors present in the ambient stimuli used (e.g., background, poses) may have influenced trait ratings, and upon which the discriminant function analysis then capitalized. We addressed this issue in Study 3 by again testing our hypothesis that individuals within social groups would resemble one another by examining classification accuracy in a same-age, same-gender, same-geographic area demographic. Specifically, we tested classification accuracy across six different fraternities at a single Midwestern U.S. university.

Methods

Stimuli. Publicly available fraternity composites, displaying all fraternity members in fairly standardized poses and attire (Figure 1), were collected online from a single university in the Midwestern U.S. to control for potential geographic variation in appearance. To control for temporal variance in appearance, all composites were taken between the years 2009-2013. In total, composites from six different fraternities were collected, with the number of members ranging from 22 to 88 (M=52.5, SD=24.4). Six was the largest number of fraternity composites available from a single university, and only six fraternities were collected and included in analysis. The majority of targets were White, but 14 (4.4%) were non-White. Individual photographs (n=315) were extracted from these composites for subsequent ratings.

Figure 1. Example fraternity composite (identifying information removed).

Participants and Procedure. 233 individuals from Amazon Mechanical Turk rated these photographs in exchange for payment. Targets from all fraternities were presented in random order and rated in a manner identical to Studies 1 and 2, with the exception that trait ratings of

warmth and competence were also collected to increase potential classification accuracy. In addition, ratings of youthfulness were not collected, as all targets were university students and approximately the same age. Separate groups of participants rated each face on these traits. Using the same exclusion criteria as in Studies 1 and 2, data from 198 participants (102 female, Age M=36.8, SD=22.9) were included in the final analysis, with each photograph rated by a minimum of 38 participants on each trait. Again, ratings for each target were averaged, and the target photograph operated as the unit of analysis. All ratings were submitted to a discriminant function analysis.

Results

Cross-validation with the hold-out approach supported our hypotheses. 52.5% of the individuals in the training set were accurately classified into their fraternities. Applying this model to the test set, 40.0% of the individuals were accurately categorized into their fraternities. Because there were far fewer fraternities (n=6) than the friendship groups (n=27) in Studies 1 and 2, expected accuracy due to chance in Study 3 was approximately 16.7%. The MCC and Q approaches again tested whether 40.0% was significantly greater than the 16.7% expected by chance. Both approaches (accuracy .400 > MCC critical .352; Q = 62.72, p <.0001) indicated that the model successfully generalized beyond the training sample. Leave-one-out results were consistent with this conclusion, with 42.8% of targets accurately classified (Figure 2). These results therefore replicated Studies 1 and 2 but across targets from a much more similar demographic, again indicating that individuals within these social groups elicit social perceptions more similar to their group members, as compared to individuals in other groups.
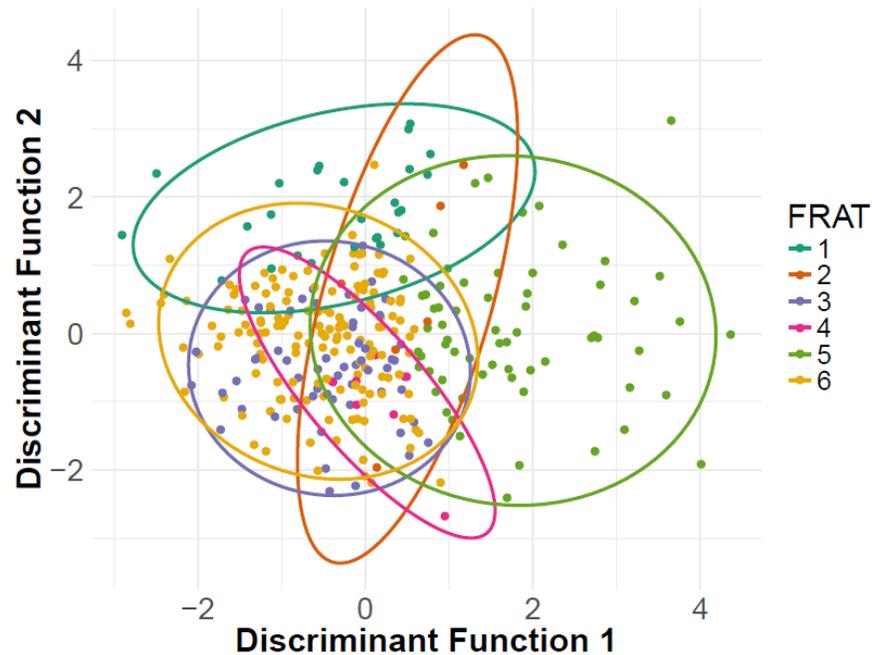
Figure 2. Discriminant function analysis results from Study 3. Individuals within each group are represented by points along the first two discriminant functions accounting for the greatest percentages of variance (x- and y-axis), though four discriminant functions were created in total (see Table 2). Ovals represent 95% confidence ellipses.

Four discriminant functions were derived from this analysis, with the first explaining a

particularly large percentage of variance (67.5%). The composition of the functions in Table 2

help to interpret Figure 2.

Traits	Discriminant Functions			
	1	2	3	4
Attractive	1.613	0.271	-0.443	0.343
Intelligent	0.12	0.358	0.582	1.291
Strong	-0.654	-0.768	0.648	0.546
Warm	-0.255	0.559	-0.357	-0.081
Competent	-0.412	0.041	0.564	-1.608
% Variance	67.5	18.7	10.2	3.6

Table 2. Composition of the four functions created by the discriminant function analysis using the hold-out approach in Study 3.

For instance, the first function was primarily composed of positive loadings on attractiveness. Given that function 1 comprises the x-axis in Figure 2, members of fraternity 5 were rated as more attractive, on average. Members of fraternity 1 are highest on function 2, the y-axis in Figure 2. Because function 2 was primarily comprised of positive loadings of warmth and negative loadings of strength, this fraternity was on average perceived as physically weaker but friendlier than the rest of the fraternities.

Though non-White targets comprised only 4.4% of the sample, to ensure that accurate classification was possible without race-based variation, we repeated all analyses with these targets removed. Results were fully consistent with those repeated here, and a full description of these analyses are available in the Supplementary Materials.

Study 4

To this point, a limitation has been that physical resemblance has been inferred from indirect measurement of the individuals within different groups by using social perceptions of these targets. While models of person perception hold that these social perceptions arise from morphological features (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Zebrowitz et al., 2003), an alternative approach is to measure these features directly. Whereas the Facebook stimuli were too noisy to measure morphological features (i.e., faces were frequently angled, rotated, or partially obscured), the photographs from the fraternity composites were ideally standardized and amenable to measurement. Therefore, Study 4 tested our hypotheses with direct measurement of the facial features of individuals within the fraternities, examining whether their group membership could be accurately classified from their morphological facial features alone.

This approach has the advantage of eliminating the possibility that factors unconsidered and thus uncontrolled for in our stimuli are not involved in accurate classification. While these

factors might influence the subjective ratings of targets, they do not influence the direct measurements from a target's face.

Methods

    Measurement of Facial Morphology. Three-dimensional computer models of the faces of fraternity members from Study 4 were created using FaceGen Modeller (Singular Inversions, 2016). Specifically, using the PhotoFit tool, key points were demarcated on each face to guide the software in creating each model, which references real anthropomorphic parameters of the human population derived from three-dimensional laser scans of several hundred male and female faces. From these measurements, 130 orthogonal components together create the symmetric shape, asymmetric shape, and texture of the facial models. Using the software development kit for FaceGen, linear combinations of these orthogonal components were extracted for analysis. These linear combinations included both shape and texture (i.e., pigmentation) parameters. However, some of the fraternity composite photographs had low-level artifacts (e.g., blurriness, graininess) consistent across all members of that fraternity, and these features were captured by the texture parameters of the computer models during import. To ensure our results were from facial resemblance and not these artifacts, we adopted a conservative approach and the texture parameters were not included in the discriminant function analysis (though including them increases accurate classification). Thus, only 62 linear combination parameters of each face were entered into the analysis (see Supplementary Materials for complete list of parameters).

Results

    Conceptually replicating Studies 1-3 but with direct measurement of facial morphology, cross-validation with the hold-out approach supported our hypotheses that individuals within

groups shared physical resemblance. 98.0% of the individuals in the training set were accurately classified into their fraternities by a five-factor solution. While this percentage initially appears impressive, it is partially a function of the larger number of variables available to the classifier, and the high accuracy serves as a potential warning of overfitting. Thus, the critical test was whether this model would generalize to the test set. The test set, however, was additionally classified with a high degree of accuracy: 78.9% of targets were correctly classified into their fraternities. The MCC and Q approaches again tested whether this result was greater than chance. Both approaches (accuracy .789 > MCC critical .248; $Q = 415.49, p < .0001$) indicated that the model successfully generalized beyond the training sample. Results from leave-one-out were consistent with this conclusion, with accuracy at 81.9% (Figure 3). Thus, from their morphological features fraternity members were classified with a high degree of accuracy, indicating that individuals within fraternities had more similar facial features to one another than to individuals in other fraternities. See Supplementary Table 15 for discriminant functions.
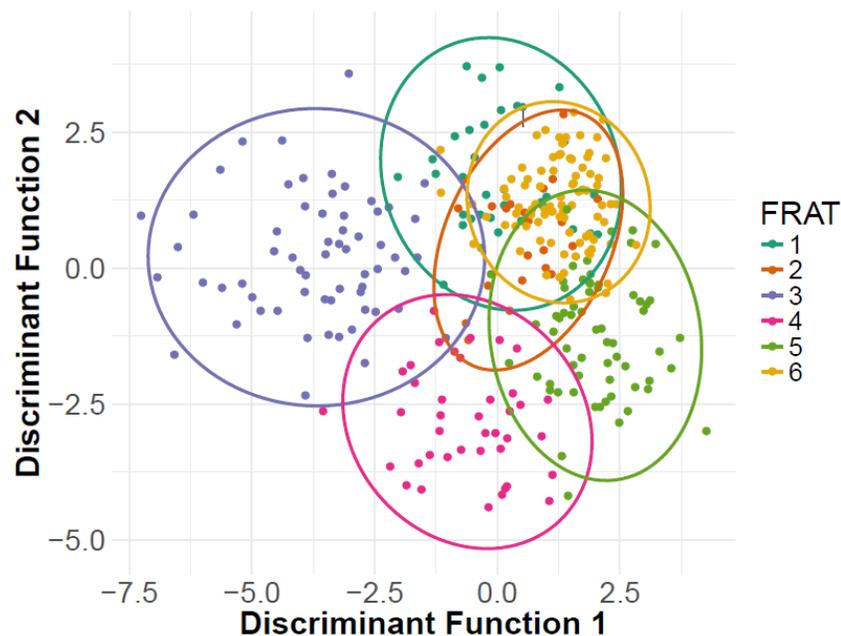
Figure 3. Discriminant function analysis results from Study 4. Individuals within each group are represented by points along the first two discriminant functions accounting for the greatest percentages of variance (x- and y-axis). Ovals represent 95% confidence ellipses.

As in Study 3, we repeated all analyses with non-White targets removed. Results were fully consistent with those reported here, and a full description of these analyses are available in the Supplementary Materials. Furthermore, the FaceGen PhotoFit tool operates best when importing faces with more neutral expressions. Due to the nature of the photographs, many of the individuals were smiling to varying extents. Accordingly, to ensure our results were not an artifact of smiling, we repeated all analyses while omitting the linear combination parameters of the chin, mouth, and jaw (i.e., those we expected to be most influenced by smiling). Results replicated the conclusions above, and a full description of these analyses are available in the Supplementary Materials. Finally, another analysis option would have been to use the orthogonal components underlying the linear combination parameters we used in the above analysis. We repeated analyses using the shape orthogonal parameters instead, and again, results were consistent with those reported here. A full description of these analyses are also available in the Supplementary Materials.

Study 5

While group membership so far has been classifiable from facial appearance, all the groups have been primarily social in nature. We have based our hypothesis on a foundation that appearance influences group acceptance decisions because individuals who resemble the group are perceived to best get along with the group. In groups in which affiliation and sociality are not a priority, however, these results may differ. For instance, groups may be primarily interested in performance or outcomes (e.g., a task force, a sports team) (Levine & Moreland, 1998). These

groups may rely less on subjective inferences of disposition arising from facial appearance, and instead seek objective information about potential future performance that might be present during the group acceptance decision making process, an effect evident in other domains (Lenz & Lawson, 2011). Therefore, groups valuing social cohesion and groups that do not might vary in how similar their group members appear. Study 5 therefore tested whether performance-based groups, who are given an abundance of information to aid group acceptance decisions, also might share physical similarity with group members, or whether this would be limited to groups valuing social cohesion. We again tested classification accuracy with both social perceptions from faces (Study 5A) and with morphological measurements (Study 5B).

Methods

   Stimuli. Like many sports, baseball is particularly rich in data regarding the objective qualifications of the players (e.g., errors, batting average). Further, though a team sport, the primary objective of the group is performance-based, to win games, with group acceptance decisions presumably made based on this objective. Therefore, baseball teams were an ideal context in which to test our hypothesis. Accordingly, photographs of the active 2015 rosters of six major league baseball teams (Blue Jays, Cubs, Dodgers, Mets, Yankees, and Red Sox) were collected from team websites (n=147). Six teams were selected to correspond with the number of fraternities in Studies 3 and 4. Only six teams were collected and included in analysis. In all photographs players were wearing team hats and these were cropped from the image.

   Participants and Procedure. For Study 5A, 438 participants recruited via Amazon's Mechanical Turk for monetary compensation rated all players' faces in a manner identical to previous studies. Participants rated targets on the same characteristics as Study 3, with the exception that we replaced intelligence with competence in this sporting context. Participants

were not informed that the faces were those of baseball players. Again, data from participants with either many consecutive identical responses or regularly responding faster than 400ms were removed, leaving 337 (211 female, $M_{age}=37.04$, $SD=12.94$) for analysis. Ratings for each target baseball player were averaged across participants, and the target functioned as the unit of analysis. All ratings were entered into a discriminant function analysis. For Study 5B, shape parameters of each digital model were created and included in a discriminant function analysis in a manner identical to Study 4.

Familiarity. Unlike the stimuli used in Studies 1-4, the baseball players in Study 5 might be familiar to some participants. Familiarity might play some role in the ratings of targets and thus contribute to classification accuracy (for subjective perceptions, it would not impact classification from the computer-generated face models). To assess and control for this possibility, a separate group of participants on Mechanical Turk ($n = 50$, $M_{age} = 40.88$, $SD = 12.61$, 35 female) were informed "You will be presented with photographs of faces of different people, some of whom you might know from media (television, movies, news, etc.). Please indicate whether faces are familiar to you." Participants responded "Yes" or "No" to every face. Familiarity was averaged for each target, with values representing the percentage of participants reporting familiarity.
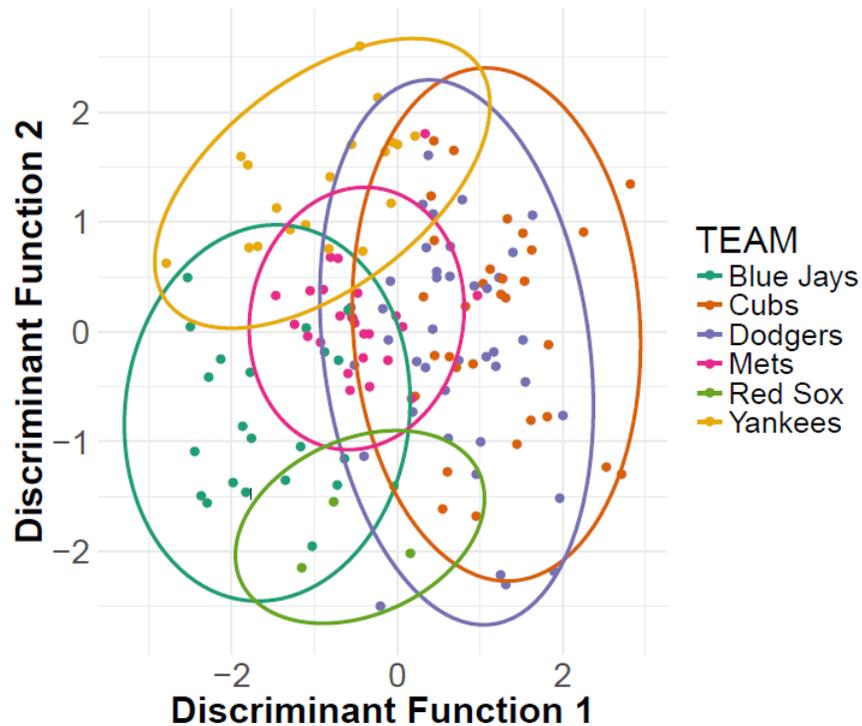
Results

Study 5A. Cross-validating the classification with the hold-out method had mixed results. 45.6% of the individuals in the training set were accurately classified into their baseball teams. The test set classified above chance with 31.5% accuracy, with chance at approximately 16.7%. While the MCC value indicated the training set did not successfully generalize (accuracy .315 < MCC critical .343), the Q test did with a significant result ($Q = 11.56$, $p = .0007$). The

leave-one-out approach also indicated the model generalized more successfully, with 29.8% of the cross-validated individuals classified correctly (Figure 4). See Table 3 for discriminant functions.

Traits	Discriminant Functions			
	1	2	3	4
Attractive	-0.422	0.915	0.538	0.45
Strong	-1.162	0.181	-0.342	-0.093
Warm	1.196	0.978	0.936	0.002
Competent	-0.082	-1.738	-0.183	0.327
Dominant	1.662	0.345	-0.007	0.712
% Variance	73	16	7.3	3.7

Table 3. Composition of the four functions created by the discriminant function analysis using the hold-out approach in Study 5A.

Figure 4. Discriminant function analysis results from Study 5A. Individuals within each group are represented by points along the first two discriminant functions accounting for the greatest percentages of variance (x- and y-axis). Ovals represent 95% confidence ellipses.

As before, and because non-White minorities comprise a greater proportion of Major League Baseball players, we repeated all analyses with non-White targets (15%) removed. Results were fully consistent with those reported here, and a full description of these analyses are available in the Supplementary Materials.

In addition, we examined to what extent familiarity with the targets might play a role in classification. On average, targets were not familiar to participants (*Mean* = .028, *SD* = .030, *Median* = .023), only 3 targets had more than 10% familiarity (2% of sample), indicating that familiarity was unlikely to play a large role in classification. Nevertheless, to statistically determine to what extent familiarity might be playing in classification, we repeated our analyses while additionally including this variable in the model. Examining the discriminant functions created from models including familiarity revealed it played a small role overall, primarily contributing to the 4th and 5th functions, which were responsible for 5.9% and .4% of the variance, respectively. See Supplementary Table 10 in the Supplementary Materials. Accordingly, for this data, we concluded that familiarity plays only a trivial role in classification accuracy.

Study 5B. Examining whether baseball players could be accurately classified from facial morphology yielded greater accuracy. Cross-validation with the hold-out approach indicated the model did generalize. 98.6% of the individuals in the training set were accurately classified into their baseball teams, but again, this is partially a function of the larger number of variables available to the classifier. The critical test of whether this model generalized to the test set indicated a seemingly high degree of accurate classification, with 49.3% of targets accurately

classified onto their team (Figure 5). Both the MCC and Q approaches indicated that this result was acceptable (accuracy .493 > MCC critical .336; $Q = 51.23$, $p < .0001$). Results from leave-one-out, with 54.6% of targets accurately classified, additionally indicated the model successfully generalized beyond the training set.

Again, we repeated all analyses with non-White targets removed. Results were fully consistent with those reported here, and a full description of these analyses are available in the Supplementary Materials. Furthermore, given the nature of the photos, many individuals were again smiling. To ensure our results were not an artifact of smiling, we repeated all analyses while omitting the linear combination parameters of the chin, mouth, and jaw (i.e., those we expected to be most influenced by smiling). The results reported above were replicated, and a full description of these analyses are available in the Supplementary Materials. Finally, we again performed analyses using the orthogonal components underlying the linear combination parameters. Results were fully consistent with those reported here. A full description of these analyses are also available in the Supplementary Materials.

Across Studies 5A and 5B and the numerous analyses in the Supplementary Materials, results consistently indicated that individuals within groups looked similar. One of the 18 tests, the MCC test from the hold-out approach of Study 5A reported above, did not support this conclusion. Taken together, we conclude that, despite the abundance of other information with which to select group members in this context, facial appearance remained an important predictor. While this result provides support for our overall hypothesis, it ran counter to our initial expectations. We further consider reasons why facial appearance may remain predictive of group acceptance even in objective qualification-rich contexts in the General Discussion. See Supplementary Table 16 for discriminant functions.
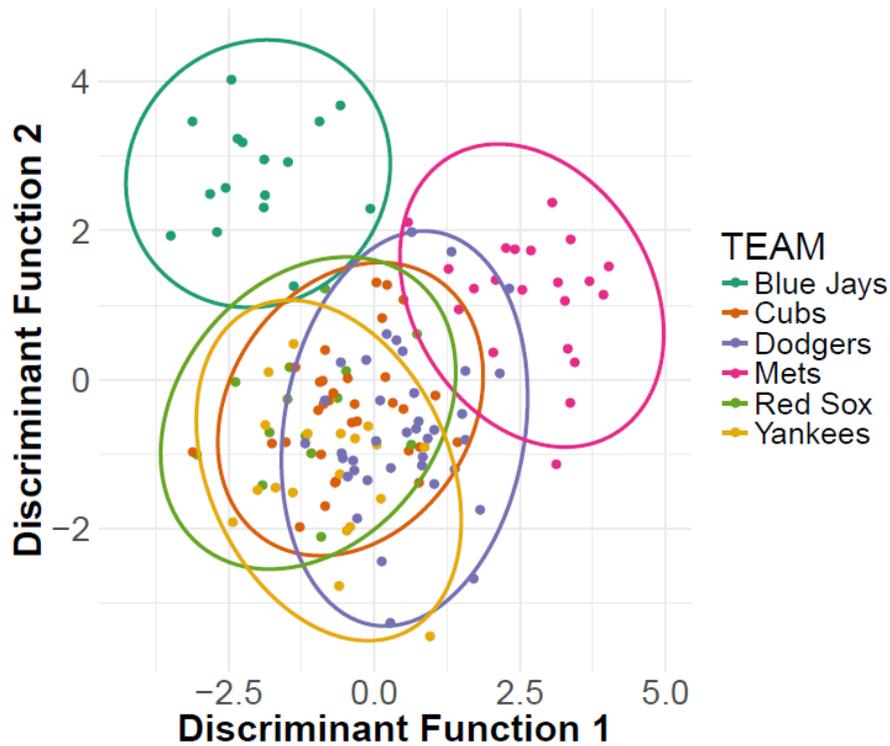
Figure 5. Discriminant function analysis results from Study 5B. Individuals within each group are represented by points along the first two discriminant functions accounting for the greatest percentages of variance (x- and y-axis). Ovals represent 95% confidence ellipses.

General Discussion

The current research tests whether the faces of individuals within groups more resemble one another than those in other groups. Cross-validated solutions from discriminant function analyses provide support for this hypothesis across six studies. Because group membership can be classified with above chance accuracy from both ratings and measurements of faces, we conclude that the faces of individuals within groups physically resemble one another.

Evidence in support of our hypothesis was found across three different types of groups: Facebook friendship groups, fraternities, and baseball teams. Facebook friendship groups and fraternities are elective and social in nature, and our theoretical focus was on these types of

groups as membership changes over time. Because membership in affiliative groups is not heritable or assigned as it is in other types of groups (e.g., race, gender, minimal groups), group boundaries are more permeable and facial appearance has the opportunity to exert an influence on membership. Specifically, since individuals use facial appearance to make inferences into others' disposition (Adams, Nelson, Soto, Hess, & Kleck, 2012; Todorov & Uleman, 2002, 2004), we theorized that facial appearance might play an important role in the formation of these groups, and indeed, we find consistent evidence for this hypothesis (but see below for discussion of alternative explanations).

Baseball teams, on the other hand, are not affiliative in the same manner. The primary goal of a baseball team is arguably to win baseball games, and while social cohesion is likely considered important by the players and management to some extent, it is the team's performance that is associated with both career success and large financial incentives. In addition, baseball management has available detailed statistical information on the performance of players, coupled with a rich history of choosing players based on these statistics (Cook, 1964). Accordingly, we had speculated that the disposition of baseball players (inferred from facial appearance) would influence group membership less than the objective performance statistics, and thus that group membership would be less accurately categorized from baseball players' facial appearance. Results generally indicate that our hypothesis was incorrect. Even the group membership of baseball players was classified well above chance from both social perceptions and direct measurement of their faces. That this result emerged despite the objective performance metrics available to baseball management may testify to the robustness of facial appearance on group membership.

One possibility for why facial appearance remained predictive of group membership in this objective information rich, baseball context involves the overall amount of variance in a candidate pool on a particular factor. To illustrate, consider an academic assistant professor search in which three candidates are invited to interview. All candidates might be excellent and fairly equal in their professional accomplishments, and thus have reduced variability in their abilities. Instead, they may vary more so on an alternative dimension, such as friendliness, that is ostensibly unrelated to the hire. In this context, friendliness might be used to determine group acceptance *from within the candidate pool*, even though the group is primarily concerned with objective ability. Accordingly, alternative characteristics conveyed by facial appearance may have determined group acceptance in this baseball context, if the abilities of the final candidates were considered relatively equal. Such a possibility would explain the current pattern of results, but evidence for this possibility cannot be determined from the available data. Future research adopting a different design might consider this possibility further. What is clear, however, is that facial appearance predicted group membership even in these performance-oriented groups, and future research might continue to determine the boundary conditions of when facial appearance is *not* associated with group membership.

That group membership can be accurately classified from faces is consistent with ecological theories of groups (Brewer, 1991; McPherson, 1983). These theories hold that groups, over time, come to occupy specialized niches in their environment that maximize their unique characteristics and values. To the extent that facial appearance is used to infer disposition, then individuals who resemble one another and value different characteristics in candidates may become more distinct over time. Hints of support for this possibility are most evident and interpretable in Study 3 examining fraternity membership. For instance, the discriminant

functions in this study help to identify the physically attractive fraternity (Frat 5) from the less attractive but friendlier fraternity (Frat 1). Individuals are attracted to affiliate with others perceived as similar (McPherson et al., 2001), and candidates who look similar may be more likely to be accepted by the group, which over time would result in groups becoming more physically distinct in facial appearance (or any other characteristic used to determine group membership).

This hypothesis, however, is the not the only one that can explain the present observed effects. An alternative, group socialization account, would be that individuals change their appearance *following* acceptance into a group. In other words, individuals might seek to maximize their social connections within the group by subtly altering their appearance. Such alterations might include hairstyle, style of dress, or even adopted poses in a photograph. This account is not necessarily independent of the mechanism posed earlier in the paper, that individuals already sharing similarity with the group are more likely to be accepted, and both might be at play in real world groups. However, it is difficult to see this potential mechanism explaining some of the present effects. Namely, some of the target photos revealed no information other than the face (i.e., the baseball photos). In these photographs, hairstyle, clothing, or other external features could not have influenced ratings, and pose is fully standardized. Similarly, Studies 4 and 5B, using measurements from the computer-generated facial models as input for analyses, would clearly not be influenced by these factors, only the actual shapes of the faces themselves. We note that these approaches cannot entirely rule out that subtle emotional expressions in the face influence these measurements, though they do decrease their likelihood. Therefore, we believe there is strong evidence that individuals are more likely to be accepted into a group the more they resemble the existing group, but the alternative that

individuals change their appearance/behaviors to better match the group following group acceptance cannot be ruled out.

Furthermore, it is plausible that our results are partially a function of the existing groups examined (i.e., Facebook friendship groups, fraternities, baseball teams). Though there is an abundance of evidence that individuals want to affiliate with similar others (McPherson et al., 2001), sometimes opposites can attract. For instance, dominant individuals report greater satisfaction with submissive, complementary partners, and *vice versa* (Dryer & Horowitz, 1997). The same might occur in some group contexts and along different traits. Individuals might be more interested in affiliating with complementary, rather than similar, others. Accordingly, our results might only generalize to groups sharing the characteristics consistently displayed by the groups observed here: medium to large, gender homogenous groups that are concerned to some extent with social affiliation. Future research using the techniques demonstrated here might examine to what extent these effects occur in groups of different types and with different goals.

The discriminant function analysis approach adopted by the current research is a data-driven technique, optimally blending whatever information available to most accurately estimate group membership. While the research questions of the present work were addressed by cross-validation classification accuracy, or *whether* group membership could be classified from faces, the composition of discriminant functions created can address *how* group membership was classified from faces. In other words, what specific social perceptions or morphological features best predict membership for different groups and in different contexts? Addressing this question thoroughly would require examining numerous types of groups in numerous contexts, and though beyond the scope of the present work, may prove a fruitful avenue of future research.

In several circumstances classification of members' faces into groups was far more accurate than initially expected. Specifically, we refer to Study 4 in which fraternity members were categorized from parameters derived from the computer-generated face models. Here, the group membership of approximately 80% of the test samples were accurately classified. Initially skeptical of this result, we performed many additional analyses reported both in the main text and Supplementary Materials which rule out potential confounds. All these additional analyses showed consistently high classification. Furthermore, we performed another simulation, classifying targets with randomly generated variables (fully reported as Supplementary Simulation 2 in the Supplementary Materials), which confirmed that such accurate classification was not an artifact of our approach using the large number of parameters from the computer-generated face models. Thus, while we have ruled out many potential confounds and alternative explanations, we urge researchers to accept this particular result tentatively until it can be confirmed by additional research.

One limitation of the present results is that we observe evidence for our hypotheses only indirectly. Though we document that the faces of individuals within groups physically resemble one another, and have hypothesized that this is due to a) individuals seeking out membership with similar-looking others and b) groups being more likely to accept similar-looking candidates, with already-existing groups no direct evidence is possible. Supplementary Simulation 1 in the Supplementary Materials demonstrates that our mechanism is viable, but to fully determine whether this process occurs in real-world groups the acceptance process of these groups would need to be observed over time, directly linking candidate appearance with acceptance. Thus, we consider the present work an initial demonstration of a phenomenon that individuals within

groups share physical resemblance, and an important but limited first step in understanding how facial appearance contributes to group membership.

Furthermore, the phenomenon demonstrated here may contribute to several psychological processes such as the perception of entitativity or the emergence of group stereotypes. To the extent that the faces of individuals within groups physically resemble one another, they may be perceived as more similar, homogenous, and thus more entitative (L. Gaertner & Schopler, 1998; Lickel et al., 2000). Similarly, when individuals within a group share physical characteristics, from which inferences into members' dispositions arise, generalizations or stereotypes about a group's physical characteristics and demeanor might more easily emerge. Furthermore, at least with regard to physical characteristics, these generalizations are more likely to be accurate for groups that share facial resemblance (Jussim, Crawford, & Rubinstein, 2015). A tendency toward perceiving outgroups as homogenous in appearance has already been well-documented (Judd & Park, 1988; Park & Judd, 1990), and the current results would only exacerbate these effects. We stress, however, that these results and conclusions are restricted to affiliative, elective groups for which membership is not heritable or assigned.

In summary, we provide the first evidence that the faces of people within affiliative groups more physically resemble one another than individuals in other groups. As such, these results have theoretical and practical implications for research examining how affiliative groups form and are maintained over time, as well as for the perception of entitativity and homogeneity in appearance of these groups.

References

Adams, R. B., Nelson, A. J., Soto, J. a., Hess, U., & Kleck, R. E. (2012). Emotion in the neutral face: A mechanism for impression formation? *Cognition & Emotion*, *26*(3), 431–441. http://doi.org/10.1080/02699931.2012.666502

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

Alvarez, L., & Jaffe, K. (2004). Narcissism guides mate selection: Humans mate assortatively, as revealed by facial resemblance, following an algorithm of "self seeking like. *Evolutionary Psychology*, *2*, 177–194.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256–274. http://doi.org/10.1037//0033-2909.111.2.256

Babyak, M. a. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*(3), 411–421. http://doi.org/10.1097/01.psy.0000127692.23278.a9

Bernstein, M. J., Young, S. G., & Hugenberg, K. (2007). The Cross-Category Effect, *18*(8), 706–712.

Berry, D. S. (1991). Accuracy in social perception: contributions of facial and vocal information. *Journal of Personality and Social Psychology*, *61*(2), 298–307. http://doi.org/10.1037//0022-3514.61.2.298

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer (Eds.), *Advances in Social Cognition* (pp. 1–36). Hillsdale, NJ, England: Lawrence Erlbaum Associates.

Brewer, M. B. (1991). The social self: On being the same and different at the same time. *Personality and Social Psychology Bulletin*.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: University of Chicago Press.

Caspi, A., & Herbener, E. S. (1990). Continuity and change: Assortative marriage and the consistency of personality in adulthood. *Journal of Personality and Social Psychology*, *58*(2), 250–258. http://doi.org/10.1037/0022-3514.58.2.250

Cikara, M., Botvinick, M. M., & Fiske, S. T. (2011). Us versus them: social identity shapes neural responses to intergroup competition and harm. *Psychological Science*, *22*(3), 306–13. http://doi.org/10.1177/0956797610397667

Cook, E. (1964). *Percentage Baseball*. MIT Press.

DeBruine, L. M. (2002). Facial resemblance enhances trust. *Proceedings. Biological Sciences / The Royal Society*, *269*(1498), 1307–12. http://doi.org/10.1098/rspb.2002.2034

Deegan, M. P., Hehman, E., Gaertner, S. L., & Dovidio, J. F. (2015). Positive Expectations Encourage Generalization From a Positive Intergroup Interaction to Outgroup Attitudes. *Personality and Social Psychology Bulletin*, *41*(1), 52–65. http://doi.org/10.1177/0146167214556240

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*(1), 62–68. http://doi.org/10.1037//0022-3514.82.1.62

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the Nature of Prejudice: Automatic and Controlled Processes. *Journal of Experimental Social Psychology*, *33*(5), 510–540. http://doi.org/10.1006/jesp.1997.1331

Dryer, D. C., & Horowitz, L. M. (1997). When do opposites attract? Interpersonal complementarity versus similarity. *Journal of Personality and Social Psychology*, *72*(3), 592–603. http://doi.org/10.1037/0022-3514.72.3.592

Efron, B., & Gong, G. (1983). A Leisurely Look at the Bootstrap , the Jackknife , and Cross-Validation Author (s): Bradley Efron and Gail Gong Source : The American Statistician , Vol . 37 , No . 1 , (Feb ., 1983), pp . 36-48 Published by : American Statistical Association Stable. *The American Statistician*, *37*(1), 36–48.

Field, A. (2009). *Discovering statistics using SPSS*. Sage.

Gaertner, L., & Schopler, J. (1998). Perceived ingroup entitativity and intergroup bias: An interconnection of self and others. *European Journal of Social Psychology*, *28*(6), 963–980. http://doi.org/10.1002/(sici)1099-0992(1998110)28:6<963::aid-ejsp905>3.0.co;2-s

Gaertner, S. L., & Dovidio, J. F. (2000). *Reducing intergroup bias: the common ingroup identity model*. Psychology Press.

Gray, K., Rand, D. G., Ert, E., Lewis, K., Hershman, S., & Norton, M. I. (2014). The emergence of "us and them" in 80 lines of code: modeling group genesis in homogeneous populations. *Psychological Science*, *25*(4), 982–90. http://doi.org/10.1177/0956797614521816

Halberstadt, J., Jackson, J. C., Bilkey, D., Jong, J., Whitehouse, H., McNaughton, C., & Zollmann, S. (2016). Incipient Social Groups: An Analysis via In-Vivo Behavioral Tracking. *Plos One*, *11*(3), e0149880. http://doi.org/10.1371/journal.pone.0149880

Hehman, E., Carpinella, C. M., Johnson, K. L., Leitner, J. B., & Freeman, J. B. (2014). Early Processing of Gendered Facial Cues Predicts the Electoral Success of Female Politicians. *Social Psychological and Personality Science*, *5*(7), 815–824. http://doi.org/10.1177/1948550614534701

Hehman, E., Leitner, J. B., Deegan, M. P., & Gaertner, S. L. (2015). Picking Teams: When Dominant Facial Structure is Preferred. *Journal of Experimental Social Psychology*, *59*, 51–59. http://doi.org/10.1016/j.jesp.2015.03.007

Hehman, E., Leitner, J. B., & Freeman, J. B. (2014). The Face-Time Continuum: Lifespan Changes in Facial Width-to-Height Ratio Impact Aging-Associated Perceptions. *Personality and Social Psychology Bulletin*. http://doi.org/10.1177/0146167214552791

Hehman, E., Mania, E. W., & Gaertner, S. L. (2010). Where the division lies: Common ingroup identity moderates the cross-race facial-recognition effect. *Journal of Experimental Social Psychology*, *46*(2), 445–448. http://doi.org/10.1016/j.jesp.2009.11.008

Hewstone, M., & Swart, H. (2011). Fifty-odd years of inter-group contact: From hypothesis to integrated theory. *British Journal of Social Psychology*, *50*(3), 374–386. http://doi.org/10.1111/j.2044-8309.2011.02047.x

Hornsey, M. J., & Hogg, M. a. (2000). Intergroup Similarity and Subgroup Relations: Some Implications for Assimilation. *Personality and Social Psychology Bulletin*, *26*(8), 948–958. http://doi.org/10.1177/01461672002610005

Inversions, S. (2016). FaceGenModeller. Toronto.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–23. http://doi.org/10.1016/j.cognition.2011.08.001

Judd, C. M., & Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. *Journal of Personality and Social Psychology*, *54*(5), 778–788. http://doi.org/10.1037/0022-3514.54.5.778

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (In)Accuracy in Perceptions of Groups and Individuals. *Current Directions in Psychological Science*, *24*(6), 490–497. http://doi.org/10.1177/0963721415605257

Kenny, D. a. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review*, *98*(2), 155–163. http://doi.org/10.1037//0033-295X.98.2.155

Kenny, D. a., & Albright, L. (1987). Accuracy in interpersonal perception: a social relations analysis. *Psychological Bulletin*, (3), 390–402.

Kubota, J. T., Banaji, M. R., & Phelps, E. a. (2012). The neuroscience of race. *Nature Neuroscience*, *15*(7), 940–8. http://doi.org/10.1038/nn.3136

Lee, R. Y., & Bond, M. H. (1998). Personality and roommate friendship in Chinese culture. *Asian Journal Of Social Psychology*, *1*(2), 179–190. http://doi.org/10.1111/1467-839X.00012

Lenz, G. S., & Lawson, C. (2011). Looking the Part: Television leads less informed citizens to vote based on candidates' appearance. *American Journal of Political Science*, *55*(3), 574–589.

Levine, J. M., & Moreland, R. L. (1998). Small groups. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (2nd ed., Vol. 2, pp. 415–469). Oxford University Press.

Lickel, B., Hamilton, D. L., Wieczorkowska, G., Lewis, a, Sherman, S. J., & Uhles, a N. (2000). Varieties of groups and the perception of group entitativity. *Journal of Personality and Social Psychology*, *78*(2), 223–246. http://doi.org/10.1037/0022-3514.78.2.223

Mackinnon, S. P., Jordan, C. H., & Wilson, A. E. (2011). Birds of a feather sit together: physical similarity predicts seating choice. *Personality and Social Psychology Bulletin*, *37*, 879–892. http://doi.org/10.1177/0146167211402094

McPherson, M. (1983). An ecology of affiliation. *American Sociological Review*, *48*(4), 519–532.

McPherson, M., Smith-Lovin, L., & Cook, J. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(2001), 415–444.

Morrison, donald g. (1969). On the interpretation of discriminant analysis. *Journal of Marketing Research*, *6*(2), 156–163.

Moskowitz, D. S. (1990). Convergence of self-reports and independent observers: Dominance and friendliness. *Journal of Personality and Social Psychology*, *58*(6), 1096–1106.

http://doi.org/10.1037//0022-3514.58.6.1096

Mummendey, A., & Wenzel, M. (1999). Social Discrimination and Tolerance in Intergroup Relations: Reactions to Intergroup Difference. *Personality and Social Psychology Review*, *3*(2), 158–174. http://doi.org/10.1207/s15327957pspr0302

Neuberg, S. L., & Fiske, S. T. (1987). Motivational influences on impression formation: outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology*, *53*(3), 431–44.

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*(32), 11087–92. http://doi.org/10.1073/pnas.0805664105

Park, B., & Judd, C. M. (1990). Measures and models of perceived group variability. *Journal of Personality and Social Psychology*, *59*(2), 173–191. http://doi.org/10.1037/0022-3514.59.2.173

Ratner, K. G., & Amodio, D. M. (2013). Seeing "us vs. them": Minimal group effects on the neural encoding of faces. *Journal of Experimental Social Psychology*, *49*(2), 298–301. http://doi.org/10.1016/j.jesp.2012.10.017

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: behavioral and neural correlates. *Journal of Personality and Social Psychology*, *104*(3), 409–26. http://doi.org/10.1037/a0031050

Stolier, R. M., & Freeman, J. B. (2016). Neural pattern similarity reveals the inherent intersection of social categories. *Nature Neuroscience*, *19*(November 2015), 1–8. http://doi.org/10.1038/nn.4296

Sutherland, C. A. M., Oldmeadow, J. a., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118. http://doi.org/10.1016/j.cognition.2012.12.001

Sutherland, C. A. M., Young, A. W., Mootz, C. A., & Oldmeadow, J. A. (2014). Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated. *British Journal of Psychology (London, England : 1953)*, 1–23. http://doi.org/10.1111/bjop.12085

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science (New York, N.Y.)*, *308*(5728), 1623–6. http://doi.org/10.1126/science.1110589

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2014). Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology*, (August 2014), 1–27. http://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., & Porter, J. M. (2014). Misleading First Impressions: Different for Different Facial Images of the Same Person. *Psychological Science*, (May). http://doi.org/10.1177/0956797614532474

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors's faces: Evidence from a false recognition paradigm. *Journal of Personality and Social*, *83*(5), 1051–1065. http://doi.org/10.1037//0022-3514.83.5.1051

Todorov, A., & Uleman, J. S. (2004). The person reference process in spontaneous trait inferences. *Journal of Personality and Social Psychology*, *87*(4), 482–93. http://doi.org/10.1037/0022-3514.87.4.482

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

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: the SHINE toolbox. *Behavior Research Methods*, *42*(3), 671–684. http://doi.org/10.3758/BRM.42.3.671

Wilson, J. P., & Rule, N. O. (2015). Facial Trustworthiness Predicts Extreme Criminal-Sentencing Outcomes. *Psychological Science*. http://doi.org/10.1177/0956797615590992

Zebrowitz, L. A., Fellous, J.-M., Mignault, A., & Andreoletti, C. (2003). Trait impressions as overgeneralized responses to adaptively significant facial qualities: evidence from connectionist modeling. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc*, *7*(3), 194–215.

Zebrowitz, L. A., & McDonald, S. (1991). The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior*, *15*(6), 603–623.

Footnotes

[1]Press' Q statistic calculated as $[N - (c * k)]^2 / [N (k - 1)]$ where N=size of test sample, c=number targets correctly classified, and k=number of groups.

[2]We use "approximately" because not all group sizes were equal. Prior probabilities of accurate classification for each group was in actuality based on the proportion of individuals in a group relative to the total sample.