Running head: FACES OF LEADERSHIP

The Many (Distinctive) Faces of Leadership: Inferring Leadership Domain from Facial Appearance

Christopher Y. Olivola [a]
olivola@cmu.edu (corresponding author)
(+001) 412-660-3508

Dawn L. Eubanks [b]
dawn.eubanks@wbs.ac.uk
+44(0) 24765 24985

Jeffrey B. Lovelace [c]
jeffrey.lovelace@usma.edu
845-938-5002

[a] Tepper School of Business Carnegie Mellon University Posner Hall #255-B 5000 Forbes Ave. Pittsburgh, PA 15213, USA	[b] University of Warwick Behavioural Science Group Warwick Business School Coventry, Warwickshire CV4 7AL, UK	[c] United States Military Academy Department of Behavioral Sciences and Leadership United States Military Academy West Point, NY 10996, USA

Disclaimer:
Any opinion expressed in whole or in part are those of the authors and do not represent those, nor reflect the official policy, of the Office of Naval Research, Department of the Army, or the US Department of Defense.

Acknowledgments:
This research was supported by a Newton International Fellowship from the Royal Society and The British Academy (to C.Y.O.). In addition, we gratefully acknowledge the financial assistance of the Behavioural Science Global Research Priorities program at the University of Warwick. We thank Vanshika Agarwala, Tyson Hayes, Alexander Mushore, Maria Okun, and Aditi Somani for providing research assistance. Finally, we would like to thank Panu Poutvaara and three anonymous reviewers for their helpful comments on earlier drafts of the manuscript.

Abstract

Previous research has shown that people form impressions of potential leaders from their faces and that certain facial features predict success in reaching prestigious leadership positions. However, much less is known about the accuracy or meta-accuracy of face-based leadership inferences. Here we examine a simple, but important, question: Can leadership domain be inferred from faces? We find that human judges can identify business, military, and sports leaders (but not political leaders) from their faces with above-chance accuracy. However, people are surprisingly bad at evaluating their own performance on this judgment task: We find no relationship between how well judges think they performed and their actual accuracy levels. In a follow-up study, we identify several basic dimensions of evaluation that correlate with face-based judgments of leadership domain, as well as those that predict actual leadership domain. We discuss the implications of our results for leadership perception and selection.

Keywords: business leadership; military leadership; sports leadership; political leadership; nonverbal behavior; social perception; implicit leadership theories

The Many (Distinctive) Faces of Leadership:
Inferring Leadership Domain from Facial Appearance

Introduction

Understanding the factors that predict leader selection is clearly important: A leader influences the achievements of his/her organization and, by extension, the wellbeing of its members and all those who benefit (or suffer) from the organization's output. Therefore, organizations and their members should have strong incentives to identify and select effective leaders within their domain, namely by relying on objective indicators of leadership quality. Yet, the human mind often relies on superficial cues to form judgments or make decisions, and the choice of which leader to select is no exception: A large and growing literature shows that facial appearances predict success in reaching prestigious leadership positions (Antonakis & Jacquart, 2013; Olivola & Todorov, 2010a). In the domain of politics, numerous studies have found that more competent-looking political candidates garner larger vote shares (e.g., Antonakis & Dalgas, 2009, Ballew & Todorov, 2007; Poutvaara et al., 2009; for a review of this literature, see Olivola & Todorov, 2010a). Voters also seem to favor more attractive candidates (Berggren, Jordahl, & Poutvaara, 2010; Efran & Patterson, 1974) and those who look stereotypically like members of their preferred political party (Olivola, Sussman, Tsetsos, Kang, & Todorov, 2012). Similarly, in the domain of business, studies have found that CEOs who possess certain facial features command higher salaries and are hired by more successful companies (Graham, Harvey, & Puri, 2013; Harms, Han, & Chen, 2012; Livingston & Pearce, 2009; Pfann, Biddle, Hamermesh, & Bosman, 2000;

Rule & Ambady, 2008; Wong, Ormiston, & Haselhuhn, 2011). And in the military domain, facial dominance was found to predict military rank (Mazur, Mazur, & Keating, 1984; Muller & Mazur, 1996; 1997; although see Loehr & O'Hara, 2013, for evidence that facial morphological correlates of dominance and aggression negatively predict military rank). In sum, there is ample research demonstrating associations, within several leadership domains (politics, business, military, etc.), between certain facial characteristics and success[1]. Leaders in a particular domain (e.g., politics) who possess the "right" facial features (e.g., a competent-looking face) tend to be more successful within that domain (e.g., receive more votes) than other (potential) leaders in the same domain who do not possess those features, ceteris paribus.

While the relationship between facial appearance and *success within* leadership domains is now well established, much less is known about the relationship between facial appearance and *selection into* particular leadership domains. That is, are certain (visible) facial features associated with being a leader in one domain rather than another? Or, to put it differently, can people discriminate between leaders in one domain (e.g., military leaders) and those in another (e.g., business leaders), just by looking at their faces? This question is important: If leaders in a particular domain share facial features that distinguish them from leaders in other domains, this suggests that domain-specific facial stereotypes may also influence the leadership selection process, above-and-beyond

[1] Here, we use the words "success", "successful", and "leadership success" to refer to the likelihood that a person is selected to a prestigious leadership position. To clarify, we are not referring to that person's leadership abilities and qualifications, nor to any successes he/she brings to their organization. The traits that make someone a popular candidate for a leadership position may well be different from those that make him/her a competent leader, once in that position. Our use of "success" (and its extensions) refers to the former (popularity), not the latter (competence).

facial cues that are broadly associated with leadership success across several domains (e.g., attractiveness and facial competence). Identifying such domain-specific facial stereotypes would therefore add a new "layer" to the role of face-based inferences in leadership selection.

This paper contributes to this important question in four ways. First, we determine whether people can accurately judge leadership domain from facial cues. To do so, we presented judges with the faces of leaders drawn from four different domains (business, military, politics, and sports) and asked them to infer which domain these leaders belong to. While there is an extensive literature on the (in)accuracy of appearance-based first impressions (e.g., Hassin & Trope, 2000; Olivola & Todorov, 2010b; Zebrowitz & Collins, 1997; Zebrowitz & Montepare, 2008), only a small fraction of these studies have specifically looked at judgments about leaders. Moreover, these studies have either examined the ability of judges to infer specific characteristics about leaders within a particular domain, such as their political orientation (e.g., Carpinella & Johnson, 2013; Jahoda, 1954; Olivola & Todorov, 2010b; Olivola et al. 2012; Wänke, Samochowiec, & Landwehr, 2012), or their ability to determine whether or not someone is a leader (Cherulnik, Turns, & Wilderman, 1990)[2]. We know of no studies that have asked participants to infer *which* domain a leader belongs to, solely from facial cues.

Second, we examine whether some leadership categories (military leaders, business leaders, etc.) are more easily identified (from facial cues) than others. In particular, we compared the accuracy of face-based leadership inferences across different

[2] There is also evidence that people can accurately infer a target's relative organizational status (Barnes & Sternberg, 1989; Schmid Mast & Hall, 2004) and behavioral indicators of dominance (Kalma, 1991), from facial photos.

leadership domains. Studies comparing face-based inferences across domains (e.g., Hassin & Trope, 2000; Olivola & Todorov, 2010b) have found that these judgments vary considerably in their accuracy levels. We might therefore expect that some leadership category inferences will be more accurate than others. In particular, it would be interesting to see whether leaders who are elected by the general population (e.g., U.S. state Governors) have more or less distinct faces than those who are selected by a smaller group of expert members within their domain (e.g., U.S. Army Generals). On one hand, we might predict that experts, being more knowledgeable (about their respective domains), would be less influenced by superficial appearance cues than most voters (Lenz & Lawson, 2011). On the other hand, since elite members of the same organization tend to be more like-minded than the general voting population, they may be more likely to share common (but possibly erroneous) stereotypes about what good leaders in their domain look like, and therefore to select leaders who possess certain, distinctive facial features. We return to this question, below, after we present the four leadership categories in our study.

Third, we assess the *meta*-accuracy of face-based leadership judgments --how well people can evaluate their own ability to draw (correct) inferences from facial stimuli. Specifically, we asked our participant-judges to report their confidence in each judgment and to estimate their overall accuracy. We then compared these estimates with their actual likelihoods of correctly inferring leadership category. Research on the validity of face-based inferences has focused, almost exclusively, on the narrow question of accuracy (see Olivola & Todorov, 2010b for a critical discussion of this issue). In contrast, much less attention has been paid to the correspondence (if any) between the

confidence that people hold in their face-based judgments (subjective accuracy) and their actual likelihood of being correct (objective accuracy). Yet meta-accuracy is an essential component of judgment validity since it determines whether (and when) one relies on appearances to form impressions: Regardless of their actual (i.e., objective) accuracy-levels, individuals who doubt their ability to draw useful inferences from faces are unlikely to deliberately rely on these judgments (and they risk ignoring a potentially useful social cue), whereas those who trust their first impressions are more likely to do so (and they risk giving these inferences too much weight)[3]. Consequently, the relative weight that individuals place on their first impressions of leaders can impact organizational dynamics, including a leader's ability to exert influence (we return to this point in the General Discussion). Therefore, an important goal for researchers should be to understand, not just whether human judges can (on average) draw accurate inferences from facial cues, but also the extent to which people recognize whether (as a general rule) and when (depending on the situation) they should rely on these inferences or refrain from doing so. Those few studies that did compare the accuracy and confidence associated with first impressions tended to find that judges were poorly calibrated in their self-evaluations (Ames, Kammrath, Suppes, & Bolger, 2010; Hassin & Trope, 2000). We might therefore predict low levels of meta-accuracy in leadership category inferences. On the other hand, given the sizeable stakes involved in selecting or interacting with leaders –in particular, the high costs of relying on invalid cues and/or failing to rely on valid cues

[3] Face-based judgments may also be partly spontaneous and perhaps difficult to control (Olivola & Todorov, 2010a; Stewart et al., 2012; Todorov, 2012). Therefore, even individuals who would rather avoid being influenced by appearances may be inadvertently affected, to some extent, by facial cues.

when evaluating a (potential) leader–, we might expect judges to be cognizant of their ability (or lack thereof) to infer leadership characteristics from appearances.

Finally, we attempt to identify the basic dimensions of evaluation (e.g., how competent a leader looks) that correlate with face-based inferences of leadership domain. That is, we examine which dimensions of evaluation might potentially underlie inferences concerning leadership domain, as well as those that may actually "give away" a leader's domain. In particular, we had the leaders' faces rated on a variety of personality traits and physical characteristics, and we correlated these ratings with guesses about leadership domain (subjective categorizations), as well as the actual domains of the leaders in our sample (objective categories). To the extent that a particular dimension of evaluation (e.g., facial competence) *simultaneously* distinguishes (i) leaders who look (stereotypically) like they belong to their domain from those (in the same domain) who do not *and* (ii) leaders in one domain from those in another, this judgment variable might underlie and contribute to the perception of leadership domain. For example, if business leaders who are easily identified (by their facial features) have very competent-looking faces compared to their less identifiable peers, *and also* compared to other types of leaders (e.g., sports leaders), this would suggest that people associate facial competence with business leadership (and vice-versa). By contrast, to the extent that a particular evaluation dimension objectively and reliably distinguishes types of leaders, *regardless* of how stereotypic they look, this variable might be an accurate predictor of leadership domain.

Study 1: Evaluating the Accuracy and Meta-Accuracy of Face-Based Leadership Domain Inferences

We recruited a large (non-student) sample of (mainly) British participants and presented them with photos of leaders in the U.S. In particular, they were shown pairs of faces drawn from two (of the four) leadership categories (e.g., business leaders and military leaders). On each trial, they were asked to judge which face belongs to a target leadership category (e.g., to identify the business leader in each pair). They also reported their confidence in each judgment. Finally, after each block of trials, they estimated how well they performed. Every participant completed two blocks of trials: one with faces drawn from two of the four leadership categories (e.g., business leaders and military leaders) and a second block with faces drawn from the remaining two domains (e.g., political leaders and sports leaders). We counterbalanced leadership pairings, target categories, and block orderings across participants.

Methods

Participants

Participants were recruited from Maximiles-UK (www.maximiles.co.uk), a British Internet service in which members earn points by completing surveys, which they can then use to purchase various consumer products (see Reimers, 2009, for additional details). Our initial sample consisted of 778 participants (97% from the U.K.). Prior to analyses, we discarded the responses of participants who either failed to complete the study, spent less than 10 minutes completing the entire study (fewer than 4% did so),

provided fewer than 70 useable trials[4] (out of 86), had an IP address that was identical to another participant's (indicating repeat survey taking), reported having previously completed the study, and/or failed any of the three 'catch' questions designed to gauge task engagement (see below). In addition, since the leaders in our study were predominantly from Western, Anglo-Saxon backgrounds, we only considered responses from participants residing in Western, Anglo-Saxon countries (99% of our initial sample). Our final sample consisted of 614 participants (44% male; Age: Range = 18-81 years, Median = 49, M = 48.29, SD = 13.39).

Leader categories & facial stimuli

Looking back at previous leadership research, certain types of leaders are frequently studied. These include business leaders, military leaders, political leaders, sports leaders, and leaders of social movements (Eubanks et al., 2010; Hunter, Cushenbery, Thoroughgood, Johnson, & Ligon, 2011; Ligon, Hunter, & Mumford, 2008; Mumford et al., 2007). In the current study, we focused on the first four categories for the following reasons: (i) there is significant precedence for studying leadership in these domains; (ii) they represent distinct categories of leadership; (iii) these individuals clearly engage in leadership activities; (iv) most leaders in these domains are not recognizable to the average person; (v) photos of these leaders are readily available; (vi) the majority of these leaders (in the U.S.) belong to the same broad demographic category (middle-aged and

[4] We discarded trials in which participants recognized one (or both) of the faces. In addition, a programming error led to a small proportion of photos (< 0.3%) being shown more than once to the same participants (these trials were also discarded). Finally, we discarded data from a few participants (n = 15) who saw the same photo presented three or more times (due to the programming error).

older Caucasian males), so one cannot merely rely on obvious cues, such as age, ethnicity, or gender, to distinguish them. The domain of social movements was not used for this study, primarily because of the widespread recognition of these individuals, their limited number, and the fact that many are not middle-aged Caucasian men (making it trivially easy for people to distinguish them from the other leadership categories).

Within each domain, we selected leaders at the top of their field, responsible for the functioning of large-scale organizations. The business leaders in our study were the chief executive officers (CEOs) of some of the 500 largest U.S. companies (http://money.cnn.com/magazines/fortune/fortune500/). CEOs are probably one of the most commonly studied leaders (e.g., Bass & Bass, 2008; Waldman, Ramirez, House, & Puranam, 2001). The military leaders in our study were decorated 4-Star Generals (some retired) on active duty status between 2007-2012, 3-Star Lieutenant Generals on active duty status in 2012, and 2-Star Major Generals in the United States Army on active duty status in 2012. Military leaders are frequently studied to improve our understanding of leadership, particularly in crisis situations (Ligon, Harris, & Hunter, 2012). The political leaders in our study were U.S. state Governors elected to office between 1996 and 2006. It is common to use political leaders for studies of leadership, particularly to understand leadership in crisis or high-pressure situations (Davis & Gardner, 2012; Eubanks et al., 2010). In contrast to other U.S. political leaders, such as Senators or Representatives, Governors are responsible for an entire state, yet they are less likely to be recognized than U.S. presidents. Finally, the sports leaders in our study were professional (NFL) and college (American) football coaches in the U.S. Specifically, we used the list of coaches studied by Hunter, Cushenbery, Thoroughgood, Johnson, and Ligon (2011). Football

coaches have considerable responsibility for the continued success of the multi-million dollar American football industry (and, in particular, the revenue and popularity of their teams). The study of sports leaders is well established (e.g., Day, Sin, & Chen, 2004; Garland & Barry, 1990; Giambatista, 2004), and sports teams are a common and accepted source for studying leadership and managerial processes (Avery, Tonidandel, Griffith & Quinones, 2003).

The leaders in our sample share some important characteristics: they are all powerful and influential individuals who make highly consequential decisions on a regularly basis. At the same time, they differ in theoretically important ways. CEOs, U.S. Army Generals, and football coaches are typically selected by small, internal groups of individuals with direct ties to their organization. Often, these individuals are peers who have had (and will continue to have) direct and regular contact with their leaders. In contrast, U.S. state Governors are elected by larger and more diverse populations of individuals. Moreover, most voters have little in common, and will never interact, with their state leaders. These differences in the leadership selection process could influence the extent to which leaders within a given domain share certain stereotypical facial features (e.g., those thought to be associated with competence in that domain). On one hand, we might predict that small groups of "insiders" would be more knowledgeable about the factors that predict leadership quality within their respective domain, and would therefore see their leadership choices being less influenced by superficial facial cues, compared to the general voting population. On the other hand, we might expect that state Governors, who are selected by large, diverse groups of individuals with conflicting preferences, beliefs, and stereotypes, will be less likely to possess common, distinctive

facial features, whereas leaders (such as U.S. Army Generals, CEOs, and football coaches) who are selected by small, uniform groups of individuals with strongly aligned preferences, beliefs, and stereotypes, will be more likely to possess common, distinctive facial features. The fact that gubernatorial elections are explicitly competitive, with rival candidates representing opposing ideologies and views, may also decrease the facial uniformity of these political leaders. Indeed, research shows that Republicans and Democrats differ, to some extent, regarding the facial features they associate with good leadership (Olivola et al., 2012). The nature of the selection pressure could thus help determine the extent to which the faces of leaders in a given domain conform to a particular stereotype. This, in turn, might influence whether (and when) human judges can accurately infer leadership category from facial cues.

In order to control for leader gender and ethnicity (which may influence face-based inferences), and since the vast majority of leaders in all four categories were Caucasian males, we excluded women and ethnic minorities (i.e., non-Caucasians) from our sample of leaders. We also excluded highly recognizable leaders (e.g., Ralph Lauren, David Petraeus) and those for whom we could not obtain a headshot of sufficient quality (e.g., because their photo was too grainy, they were not facing the camera, etc.). Our final sample of leaders consisted of 325 business CEOs, 64 U.S. Army Generals, 66 state Governors, and 43 football coaches. These leaders were all Caucasian males (as explained above) and their ages (at the time the photos were taken) generally ranged from 40 to 70 years.

We obtained headshots of the business CEOs, U.S. Army Generals, and football coaches from publically accessible Internet sources. The photos of U.S. Governors were

drawn from an existing set of stimuli used in previous studies of gubernatorial elections (see Ballew & Todorov, 2007; Olivola et al., 2012; Olivola & Todorov, 2010a); however, we only selected the winning candidates (i.e., the ones who actually went on to become Governors). We then standardized all of these photos using the following four-step procedure: (i) by converting them to black-and-white (i.e., grayscale), (ii) by cropping each headshot to remove everything outside the main contour of the face (i.e., excluding ears, hair, and neck), (iii) by resizing them to be approximately the same dimension (without distorting natural variations in facial width and height), and finally (iv) by placing them on a black background.

Experimental Design

The experiment consisted of two blocks of 43 trials, with each block pitting two leadership categories against each other so that all four categories were presented to each participant. In other words, the first block presented two of the four leadership categories (e.g., military leaders vs. business leaders) and the second block presented the remaining two categories (e.g., political leaders vs. sports leaders). Thus, each participant completed two mutually exclusive judgment tasks (one in each block). There are six possible ways to pair four leadership categories and each pair can come first (in block 1) or second (in block 2), leading to 12 combinations of pairs and block orderings. In addition, within each pair, we varied the specific leadership category (the "target category") that participants were trying to identify. For example, among participants assigned to a block of trials pitting political leaders against military leaders, half were instructed to identify the political leader in each trial (P-M category pair), while the other half were asked to

identify the military leader in each trial (M-P category pair). Although both tasks are formally equivalent, the particular target category that participants are assigned to may influence which facial cues they rely on (e.g., participants assigned to identify the political leaders might focus on facial attractiveness, whereas those assigned to identify the military leaders might instead focus on facial cues that seem to convey discipline). All together, our fully counterbalanced experimental design called for 24 conditions (see Table 1). Unfortunately, a programming error led to one condition being omitted (Condition 17 in Table 1). Fortunately, the two leadership category pairs (P-S [political leaders as the target category paired with sports leaders] and B-M [business leaders as the target category paired with military leaders]) that made up this condition were also presented in several other conditions (see Table 1), so our analyses and results were unlikely to have been seriously affected by this error. Each participant was randomly assigned to one of the remaining 23 conditions. The number of participants assigned to the P-S and B-M category pairs was n = 68 and n = 73, respectively. The number of participants assigned to the remaining category pairs ranged from n = 96 to n = 121.

Table 1 about here

Procedure

The entire study was conducted online, via a web-based experiment. Participants were sent an email containing a link to the study and an invitation to participate in exchange

for compensation (the Maximiles points, described above). A series of instruction screens introduced participants to the study, explaining what they would see and need to do. Participants were asked to complete the study attentively, on their own, and in one sitting (without taking breaks). They were informed that they would be shown faces of leaders drawn from four domains (business, military, politics, and sports), and that their task would be to guess which domain these leaders belonged to, just by looking at their faces. They were also informed that the experiment consisted of two separate parts (corresponding to the two blocks of trials). Additional instructions appeared before each block of trials, specifying which two leadership categories would appear. For example, participants assigned to a block of trials pitting business leaders against military leaders received the following instructions: "In each trial you will be shown pairs of adult faces. Each face belongs to a different kind of leader. One of them is a Business CEO, while the other is a Military General." For participants assigned to a block pitting sports leaders against political leaders, the labels "Sports Coach" and "Politician" were used. The instructions also specified the target category that they would be asked to identify throughout the block (e.g., "Your task is to guess, for each pair, which person is the Business CEO, by relying only on their facial photos"). Finally, participants were informed that the location of the target category (left vs. right half of the screen) would vary randomly across trials.

Figure 1 illustrates how each trial progressed. On each trial, participants were presented with the faces of two leaders, side by side: one from the target category and one from the other leadership category (as determined by treatment condition and block number). The location of the target face (whether it appeared on the left or right half of

the screen) varied randomly across trials. No other information about the leaders was provided. Participants indicated which person they thought belonged to the target category by clicking on his photo. Next, they indicated how confident they were in their judgment (on a 0-100% scale). Finally, they reported whether they recognized one or both leaders (using buttons below each photo –see Figure 1). Participants then advanced to the next trial. No feedback was provided about their performance.

Figure 1 about here

The selection and presentation order of face stimuli in each leadership category were randomized (separately for each participant) in the following way: On each trial, one leader from each category was randomly drawn (without replacement) and presented. This process was repeated until 43 trials were completed (at which point the block ended). Thus, participants saw all the sports leaders in our stimulus sample (i.e., all 43 football coaches, presented in random order) and a subset of 43 leaders from every other leadership category (randomly drawn, for each participant, from the entire pool of stimuli in that category).

Once the first block was completed, the instructions for the next block appeared, indicating the two remaining leadership categories that participants would be shown and specifying the new target category that they would be asked to identify on each trial. The general design and structure of the trials were identical in both blocks; the only

differences being the specific leadership categories presented and the targets to be identified. After completing each block, participants were asked to retrospectively estimate the proportion of trials (from 0-100%) in which they had correctly identified the target category. Before providing their estimate, they were informed that random guessing throughout the entire block would still achieve roughly 50% accuracy.

The final part of the experiment (after the second block was completed) consisted of a simple survey containing a series of demographics questions. Among other things, participants were asked whether they thought they could guess other people's characteristics by "reading" their faces (they could respond "Yes" or "No" to this question). They also reported the extent to which they had an interest in each of the four domains (politics, sports, business, and the military), using four-category ordinal scales ("Not at all", "To a limited extent", "To some extent", or "To a great extent"), which we recoded into 0-3 ratings of interest. In addition, the survey contained three 'catch' questions, designed to measure participant engagement. The first one asked participants to indicate the capital of England. The second one asked them to select the word "Shard" from a list of words (which included visual decoys such as the words "Shark" and "Sharp"). The final catch question asked them to report their date of birth (in order to see whether it was consistent with their earlier reported age). Finally, participants were asked whether they had previously completed the experiment. Most participants (82%) completed the experiment in 10-25 minutes (another 14% spent more than 25 minutes completing the study).

Analyses

Before carrying out our analyses we discarded trials in which either leader was reportedly recognized (1% of trials, on average). Furthermore, we only considered trials in which the two faces were matched in terms of their facial hair (moustache and/or beard) and whether they wore glasses. We then calculated each participant's judgment accuracy (the percentage of correct judgments), their average confidence level (across trials), and the point-biserial correlation (across trials) between their rated confidence (a continuous variable) and whether they correctly identified the target leader (a binary variable). All three variables were calculated separately for each block. To simplify our analyses, and since we used a counterbalanced design (so that each leadership category pair was roughly equally likely to appear in the first or second block), we treated each block as an independent observation. This is further supported by the fact that we found no main effect of block number, nor an interaction between block number and leadership category pair (when both variables were entered into an ANOVA), on any of these three dependent variables.

Results

Accuracy levels: Estimated vs. actual

Figure 2 shows how well participants thought they performed and how accurate they actually were at distinguishing leaders in each category pairing (along with the 95% confidence intervals for each mean). Despite being explicitly informed that even random guessing would generally yield chance-level accuracy, participants were rather pessimistic about their performance: mean accuracy estimates were all significantly below chance. And yet, across all leadership judgment tasks, participants performed

significantly better than they expected. In fact, mean accuracy levels significantly exceeded chance for most leadership category pairings. In other words, we found that participants could identify the domains that leaders belong to just by looking at their faces. In particular, they were generally able to pick out business leaders, military leaders, and sports leaders from pairs of black-and-white cropped facial photos, which is impressive given how little information these photos provided. Interestingly, this 'ability' does not extend to politicians: When participants were asked to identify political leaders their mean accuracy levels were no better than chance (ranging from 48% to 51%).

Figure 2 about here

Meta-accuracy & confidence calibration

Comparing average performance with average estimated performance showed that participants consistently underestimated their ability to distinguish leader types using facial cues alone. Yet this general (downward) bias in estimated performance says little about the extent to which the perceived ability to infer leadership category from faces *correlates* with the actual ability to do so. In particular, we examined two important questions concerning the calibration of meta-accuracy and confidence judgments: First, looking across participants, were the least accurate judges more modest about their ability to identify leader types than the most accurate ones? And second, looking within participants (and across-trials), were judges more confident about their judgments when

these were correct (vs. incorrect)? To examine the first question, we looked at the extent to which participants' accuracy levels correlated with (i) their estimated performance[5] and (ii) their overall confidence (averaged across trials). To examine the second question, we looked at the within-subject, cross-trial point-biserial correlation between judgment confidence and 'correctness' (whether the target leader was accurately identified). The results of these analyses, presented in Figure 3, reveal a strikingly weak relationship between accuracy and meta-accuracy, both within- and between-judges. In all but two leadership category pairings (Business-Military and Military-Politics), estimated accuracy did not significantly predict actual performance. Similarly, the average level of confidence that participants reported in each trial was not significantly related to their overall performance. And we found an extremely weak within-subject relationship between confidence in a judgment and the likelihood that this judgment was correct (all point-biserial correlations were below .08). A final result worth noting concerns the binary (Yes/No) question in the final (post-trials) survey ("In general, would you say that you can guess people's characteristics by "reading" their faces?"): Participants who responded "Yes" to this question were no better at identifying business leaders, political leaders, or sports leaders than those who responded "No" (all p-values > .21), and they were only marginally better at identifying military leaders: 57% vs. 55% accuracy, $t(318) = 1.90$, $p = .058$.

[5] We correlated estimated accuracy (but not confidence) with the proportion of correct judgments across *all* trials within a block, including those with recognized or non-matched faces. This was done because participants were asked to estimate their performance over the entire set of trials in each block (i.e., we did not instruct them to ignore recognized or non-matched trials).

Figure 3 about here

Participant age, gender, and domain-specific interest

Although participants were, on average, able to accurately identify three leadership categories from facial cues alone, this ability might vary with experience (i.e., age), gender (since all targets were male), and/or one's interest in the target domain. To test these possibilities, we regressed our measure of actual accuracy, perceived accuracy, and within-subject meta-accuracy onto participants' age, gender (dummy-coded), and reported interest in all four domains (represented by four 0-3 ratings). We carried out separate regressions (using robust error estimates) for each target leadership category and each dependent variable, with the six predictors always entered simultaneously. Surprisingly, there were very few significant predictors of accuracy or meta-accuracy. With regard to actual accuracy: age and reported interest in the military both *negatively* predicted the ability to identify U.S. Army Generals (B = -0.12%, p = .014 and B = -1.39%, p = .099, respectively), whereas reported interest in business positively predicted the ability to identify U.S. state Governors (B = 2.15%, p = .029) and football coaches (B = 2.64%, p < .008), but *not* CEOs. With regard to estimated accuracy: age marginally and positively predicted how accurately participants thought they were able to identify business CEOs and U.S. state Governors (B = 0.16%, p = .085 and B = 0.16%, p = .060, respectively), reported interest in a given domain (except sports) was associated with a greater perceived ability to identify leaders from that domain (B = 5.27%, p < .001; B =

2.31%, $p = .089$; $B = 4.12\%$, $p < .005$, for business, the military, and politics, respectively), reported interest in politics marginally and positively predicted how frequently participants thought they correctly identified U.S. Army Generals ($B = 2.32\%$, $p = .080$), and reported interest in business marginally and positively predicted how frequently they thought they correctly identified U.S. state Governors ($B = 2.39\%$, $p = .099$). Finally, with regard to (within-subject) meta-accuracy: age and reported interest in sports both negatively predicted meta-accuracy when U.S. football coaches were the targets ($B = -.0017$, $p = .074$ and $B = -.026$, $p = .036$, respectively); no other results were significant. In sum, these analyses suggest that age and domain-specific interest are at best weakly (and inconsistently) related to accuracy and meta-accuracy, whereas gender seems to play no role. The only relatively coherent pattern of results seems to be that interest in a domain predicts greater *perceived* ability to identify leaders from that domain (even after the task is completed), which is not especially surprising.

Discussion

Our first study shows that human judges can accurately infer several leadership domains from facial cues alone. However, that study tells us very little about the underlying impressions or facial cues that drive their judgments. It is worth reminding our readers that we did control for a number of possibly relevant cues, such as gender, ethnicity, clothing, hairstyle, facial hair, and glasses. Consequently, none of these variables can account for our results. Moreover, the leaders in our study were largely within the same age range. This leaves facial physiognomy (the shape of a person's face) and subtle facial expressions, as the most likely sources of inference. In an attempt to shed further light on

this question, we carried out a second study to identify some of the basic dimensions of evaluation that correlate with face-based leadership domain inferences and/or with actual leadership domain. Specifically, we had a new set of participants rate a subset of the leader faces from our first study on 15 "basic" dimensions (such as trustworthiness, likeability, etc.) that could plausibly underlie leadership category judgments (see Table 2 for the full list of dimensions included in our study). We selected dimensions of evaluation that have received considerable attention in previous studies of face-based leadership inference (see Antonakis & Jacquart, 2013; Olivola & Todorov, 2010a), as well as dimensions that have received less attention but seem potentially relevant, such as ambition and confidence. We then compared the mean ratings (i.e., impressions) that each leader elicited, as a function of his domain and how accurately he was categorized in our first study.

Study 2: The Evaluative Correlates of Face-Based Leadership Domain Inferences

Methods

Participants

We recruited a completely new sample of participants from the (mainly) British online population used in our previous study (Maximiles-UK; the link to the current study was only sent to members who had *not* participated in our first study). Our initial sample consisted of 1,105 participants (97% from the U.K.). Prior to analyses, we discarded the responses of participants who either failed to complete the study, spent less than 7 minutes or more than 60 minutes completing the entire study (fewer than 4% did so), had

an IP address that was identical to another participant's (indicating repeat survey taking), reported having previously completed the study, were less than 18 years old, and/or failed any of the three 'catch' questions designed to gauge task engagement (we used the catch questions from our previous study). In addition, as with our first study, we only considered responses from participants residing in Western, Anglo-Saxon countries (97% of our initial sample). Finally we discarded the data from one participant who provided a rating of '0' across all 80 trials. Our final sample consisted of 929 participants (43% male; Age: Range = 18-92 years, Median = 49, M = 48.11, SD = 14.09).

Facial stimuli

We selected a subsample of the facial stimuli used in our previous study, according to the following three-step procedure: First, we only considered leaders who did not have glasses and were clean-shaven (no moustache or beard) in their photo. Second, we only considered leaders whose photo had been presented to at least 30 participants in our previous study (to ensure that their accuracy scores were based on a sufficiently large sample of judgments). Finally, within each leadership domain, we selected the 10 most accurately identified leaders and the 10 least accurately identified leaders (with accuracy calculated from trials in which these leaders represented the target category). The average accuracy levels for the 10 *most* accurately identified leaders were: 72%, 65%, 56%, and 64%, in the domains of business, military, politics, and sports, respectively. The average accuracy levels for the 10 *least* accurately identified leaders were: 33%, 44%, 42%, and 44%, in the domains of business, military, politics, and sports, respectively. Thus, our

final sample of stimuli consisted of 80 headshots (20 per leadership domain) drawn from our previous study.

Experimental Design

The experiment consisted of a single block of 80 trials, with each trial presenting one leader. Leader presentation order was randomized for every participant. Each participant was randomly assigned to rate all 80 photos on single dimension (drawn from the 15 dimensions listed in Table 2). The number of participants assigned to each dimension ranged from n = 55 to n = 69.

Table 2 about here

Procedure

The procedure was broadly similar to the one used in our previous study, with three important exceptions. First, participants did not know that they were evaluating leaders. They were simply informed that they would "be shown a series of male faces." Second, instead of evaluating pairs of faces, participants were presented with a single face on each trial, and asked to evaluate[6] that person according to a (single) dimension of evaluation.

[6] In line with previous studies (e.g., Ballew & Todorov, 2007; Olivola & Todorov, 2010a; Willis & Todorov, 2006), we instructed participants not to spend too much time forming their evaluations. The specific instructions they received were as follows: "We ask that you please rely on your ''gut feeling'' to form your impressions, without thinking too

For example, some participants were asked to rate the attractiveness of each person, while others were instead asked to rate the trustworthiness of each target, and so on. All participants provided their ratings using a continuous sliding scale that ranged from "Not at all [dimension name]" to "Extremely [dimension name]". Unbeknownst to the participants, the scale ranged from 0 to 100. The slider was always set to start in the middle of the sliding bar (i.e., a rating of '50'), and participants had to move the slider before they could advance to the next trial. Third, participants were not asked to report their confidence or whether they recognized a face. Most participants (90%) completed the experiment in 7-25 minutes.

Analyses

We standardized ratings within participants (across leader photos), using a z-score transformation based on each participant's mean and standard deviation. We then averaged these z-scores across participants for each leader photo and for each dimension of evaluation. We thus obtained 15 average z-scores for each leader, corresponding to the 15 dimensions of interest. Table 2 presents the correlations between these 15 dimensions (each one based on n = 80 leader photos). As this table makes apparent, many of these dimensions are highly correlated. Therefore, to simplify our analyses, avoid multiple testing issues, and minimize redundancies, we carried out a series of Principle Component Analyses (PCAs) on these 15 dimensions (see Appendix A). The PCA results suggested that three distinct components capture the essence of these evaluations (only these three components had associated eigenvalues > 1): The first component, which we

---

much about each face. There are no right or wrong answers; we are simply interested in your first, immediate impression of each person."

call the "Warmth" score, reflects how anxious (reverse-coded), charismatic, confident, dominant (reverse-coded), likeable, threatening (reverse-coded), and trustworthy a leader looks. The second component, which we call the "Competence" score, reflects how ambitious, competent, conservative, and disciplined a leader looks. Critically, these first two dimensions have been shown to be basic and universal components of social evaluation (Fiske, Cuddy, & Glick, 2007; Judd, James-Hawkins, Yzerbyt, & Kashima, 2005). Finally, the "Masculinity-Maturity" score reflects how babyfaced (reverse-coded) and masculine a leader looks. These three scores were calculated for each leader by averaging the relevant ratings (making up each component). The "Competence" score, for example, was calculated by adding ratings of ambition, competence, conservatism, and discipline, then dividing the resulting total by four. Neither facial attractiveness nor facial extraversion loaded clearly and exclusively onto one of these components. We therefore consider attractiveness separately (given its importance in social evaluations) and ignore ratings of extraversion (since several studies have found that facial extraversion has weak predictive power – e.g., Olivola & Todorov, 2010a). All of our analyses, going forward, focus on the three components described above and facial attractiveness.

Results

Correlates of leadership domain facial stereotyping

To determine which dimensions of evaluation might drive people's inferences about leadership domain, we can compare how the most stereotypical looking and least stereotypical looking leaders are perceived. Figure 4 shows the average attractiveness and

facial component scores of the most accurately identified and least accurately identified leaders in each of the four domains. The most stereotypic-looking business leaders scored significantly higher on the facial "Competence" dimension than their least stereotypic-looking peers ($t(18) = 3.15$, $p < .006$, $d = 1.48$), whereas none of the other three dimensions of evaluation were associated with facial stereotyping in the business domain (all p-values > .5). The most stereotypic-looking military leaders scored significantly higher on the facial "Competence" ($t(18) = 2.54$, $p = .020$, $d = 1.20$) and "Masculinity-Maturity" ($t(18) = 3.50$, $p < .003$, $d = 1.65$) dimensions and significantly lower on the facial "Warmth" dimension ($t(18) = 3.51$, $p < .003$, $d = 1.65$), than their least stereotypic-looking peers, but they did not significantly differ in terms of attractiveness ($t(18) = 1.29$, *ns*). The most stereotypic-looking political leaders scored significantly higher on the facial "Competence" dimension than their least stereotypic-looking peers ($t(18) = 3.12$, $p < .006$, $d = 1.47$), whereas none of the other three dimensions of evaluation were associated with facial stereotyping in the domain of politics (all p-values > .2). Finally, the most stereotypic-looking sports leaders scored significantly higher on the facial "Warmth" dimension ($t(18) = 2.23$, $p = .039$, $d = 1.05$) and were judged to be more attractive ($t(18) = 2.46$, $p = .024$, $d = 1.16$) than their least stereotypic-looking peers, but they did not significantly differ in terms of the remaining dimensions (both p-values > .4).

We can also compare the evaluations of the most stereotypic-looking leaders across domains to see which dimensions people seem to rely on when they are trying to distinguish between leadership domains (using facial cues). ANOVAs comparing the most correctly identified leaders in each domain in terms of their facial component scores

revealed significant differences (between domains) on all four dimensions (all Fs > 3.62; all p-values < .03). As Figure 4 shows, the most stereotypic-looking military leaders scored lower on the "Warmth" dimension than their counterparts in business ($t(18) = 3.04$, $p < .008$, $d = 1.43$), politics ($t(18) = 4.69$, $p < .001$, $d = 2.21$), and sports ($t(18) = 2.75$, $p = .013$, $d = 1.30$). The most stereotypic-looking sports leaders scored lower on the "Competence" dimension than their counterparts in business ($t(18) = 3.61$, $p < .003$, $d = 1.70$) and politics ($t(18) = 2.30$, $p = .033$, $d = 1.09$), but they did not significantly differ from military leaders on this dimension ($t(18) = 1.72$, *ns*). The most stereotypic-looking military leaders scored higher on the "Masculinity-Maturity" dimension than their counterparts in business ($t(18) = 3.25$, $p < .005$, $d = 1.53$) and politics ($t(18) = 2.16$, $p = .044$, $d = 1.02$). Similarly, the most stereotypic-looking sports leaders scored higher on the "Masculinity-Maturity" dimension than their counterparts in business ($t(18) = 4.15$, $p < .001$, $d = 1.95$) and politics ($t(18) = 2.61$, $p = .018$, $d = 1.23$). Finally, the most stereotypic-looking military leaders were judged to be less attractive than their counterparts in business ($t(18) = 2.56$, $p = .020$, $d = 1.21$) and politics ($t(18) = 3.09$, $p < .007$, $d = 1.46$).

Taken together, these results suggest that people seem to associate military leadership with low warmth and high masculinity-maturity. They also seem to associate business and political leadership with competence, especially when compared to sports leaders.

Figure 4 about here

Facial impressions that predict leadership domain

While the previous analyses helped us identify a few dimensions of face-based inferences that correlate with (and might therefore drive) stereotyping about leadership domain, they do not tell us whether these facial stereotypes are accurate predictors of actual leadership domain. To determine which facial impressions might actually predict leadership domain, we need to compare face-based evaluations across domains without excluding leaders who were incorrectly categorized (since these leaders are, in fact, members of their respective domains). Figure 5 shows the average attractiveness and facial component scores of the leaders in each of the four domains (combining stereotypic-looking and non-stereotypic-looking leaders). ANOVAs comparing leaders in each domain in terms of their facial component scores revealed significant differences (between domains) on all four dimensions (all Fs > 2.76; all p-values < .05). Military leaders scored lower on the "Warmth" dimension than their counterparts in business ($t(38) = 2.08$, $p = .045$, $d = .67$) and politics ($t(38) = 4.23$, $p < .001$, $d = 1.37$). Similarly, sports leaders scored lower on the "Warmth" dimension than their counterparts in business ($t(38) = 1.74$, $p = .089$, $d = .57$) and politics ($t(38) = 3.88$, $p < .001$, $d = 1.26$), although the former difference was only marginally significant. Sports leaders also scored lower on the "Competence" dimension than their counterparts in business ($t(38) = 2.83$, $p < .008$, $d = .92$). In contrast, sports leaders scored higher on the "Masculinity-Maturity" dimension than their counterparts in business ($t(38) = 4.54$, $p < .001$, $d = 1.47$), the military ($t(38) = 2.59$, $p = .014$, $d = .84$), and politics ($t(38) = 3.11$, $p < .004$, $d = 1.01$). Business leaders were

judged to be more attractive than military ($t(38) = 2.49$, $p = .017$, $d = .81$) and sports leaders ($t(38) = 3.50$, $p < .002$, $d = 1.13$). Similarly, political leaders were judged to be more attractive than military ($t(38) = 3.76$, $p < .001$, $d = 1.22$) and sports leaders ($t(38) = 4.43$, $p < .001$, $d = 1.44$).

In sum, we find that several face-based impressions predict (actual) leadership category, at least for the four domains included in our studies. In particular, our results suggest that one could potentially distinguish military and sports leaders from business and political leaders by evaluating how warm and attractive they look (from their faces), since the former two types of leaders look less attractive and less warm than the latter two. Sports leaders could also be distinguished by the fact that they look less competent than business leaders, as well as more masculine and mature than the other types of leaders.

Figure 5 about here

Discussion

Our second study identified several basic dimensions of evaluation that correlate with face-based judgments of leadership domain, as well as those that actually predict leadership domain. By combining these two sets of results, we can start to speculate about the facial-evaluations that the participants in our first study might have been using to achieve their above-chance performance (when trying to infer leadership domain from

faces). The results of our second study suggest that people may have a few, somewhat accurate stereotypes about the relationship between facial features and leadership category, at least for two domains: They seem to correctly associate business leaders with "Competent"-looking faces[7]. They also seem to correctly associate military leaders with less attractive and less "Warm"-looking faces. Thus, it might be that face-based impressions of "Competence" allowed participants to identify business leaders with above-chance accuracy, while judgments of attractiveness and "Warmth" allowed them to identify military leaders with above-chance accuracy. The first hypothesis is further supported by the fact that participants were best able to identify business leaders when these were paired with sports leaders (Figure 2), who received the lowest face-based "Competence" ratings (on average). The second hypothesis is similarly supported by the fact that participants were not significantly better than chance at identifying military leaders when these were paired with sports leaders (Figure 2), who also received low ratings of attractiveness and "Warmth". On the other hand, this second hypothesis is harder to reconcile with the fact that military leaders were more accurately identified when paired with business leaders than with political leaders (Figure 2), even though the latter group received the highest ratings of attractiveness and "Warmth" (on average).

People also seem to possess misleading facial stereotypes concerning certain leadership domains: They seem to erroneously believe that military leaders have very masculine and mature-looking faces, and that political leaders have very "Competent"-looking faces. The second bias, in particular, would help explain why our participants

[7] Along similar lines, Graham et al. (2013) find that CEOs have more competent-looking faces than non-CEOs. Our results show that CEOs are distinguishable, not only from non-leaders, but also from other kinds of leaders, by their facial competence.

were so poor at identifying political leaders (Figure 2). We need to stress that these hypotheses, concerning basic face-based evaluations that may help (or harm) leadership domain inferences, are very preliminary and speculative since we do not have direct evidence connecting people's impressions to their leader category judgments.

A couple of other important caveats about these results are also worth discussing. First, the fact that certain face-based impressions correlate with (actual) leadership domain does not imply that these impressions are accurate. For example, the fact that sports leaders have less competent-looking faces than do business leaders does not imply that they are actually less competent. It may be that our participants picked up on facial cues that happen to correlate with sports vs. business leadership, and which they erroneously associate with objective competence. Alternatively, the relationship we observe between facial impressions and leadership domain could reflect a self-fulfilling prophecy dynamic (Antonakis, 2011). For example, if everyone shares the belief that facial warmth is a valid indicator of actual warmth and, moreover, that warmth is a liability in military leadership, this could bias the promotion process by favoring military leaders who happen to have more "cold" looking faces. This dynamic would lead to the top military leadership being populated by individuals who are less warm in appearance.

Second, the fact that leadership inferences correlate with evaluative judgments that are considered to be basic components of social perception, such as warmth and competence, does not imply that impressions of warmth and competence underlie or somehow drive face-based leadership-inferences. For example, when participants are asked to identify military leaders on the basis of facial features alone, it might be the case that they first evaluate facial warmth and use this cue as a basis for their leadership

categorization judgments (in which case, facial warmth evaluations would indeed drive face-based leadership inferences). But it could instead be the case that a third, common variable underlies both warmth and leadership judgments, leading to the relationship we observe. Or the relationship could be reversed: When participants are asked to evaluate a person's warmth, they first try to determine whether this person would be a good soldier or military leader, and they then use this cue to judge warmth. This would also produce the relationship we observe between facial warmth and inferences about military leadership. More generally, we caution readers against drawing strong conclusions from the results of our second study. The relationships we observed between these four basic evaluation dimensions (attractiveness, "Warmth", "Competence", and "Masculinity-Maturity") and leadership domains are suggestive, but they will need to be replicated and examined in more detail by future researchers.

General Discussion

Our results show that leaders from several different domains are distinguishable by their facial features. Specifically, we found that British participants could identify American business leaders (company CEOs), military leaders (U.S. Army Generals), and sports leaders (American football coaches), from a 'lineup' of two faces (belonging to leaders from two different domains) with above-chance accuracy. They were able to do so despite not recognizing either face and even though these leaders were drawn from another country. This latter point is noteworthy, since it suggests that facial stereotypes about business, military, and sports leaders may cross national and (sub)cultural borders. Interestingly, these same participants were *not* able to identify political leaders (U.S. state

Governors), which suggests this fourth group may not have unique, distinguishable facial features (that reveal their leadership domain) to the same extent as the other three groups. Whereas actual accuracy was generally greater than chance, meta-accuracy (the ability to properly evaluate one's own accuracy) was not. For one thing, participants consistently underestimated their performance (in contrast to some previous studies –e.g., Hassin & Trope, 2000). More surprisingly still, we found little or no relationship between how well they thought they had performed and their actual accuracy levels. Nor did we find much of a relationship (within- or between-subjects) between the level of confidence that they placed in their judgments and the likelihood that those judgments were correct.

We also found, in a second study, that several basic face-based dimensions of evaluation can predict leadership domain inferences and actual leadership categories. Our results suggest, for example, that people associate military leadership with low warmth and high masculinity-maturity, and they associate business and political leadership with competence. Our results also indicate that one might be able to distinguish military and sports leaders from business and political leaders by evaluating how warm and attractive they look (from their faces), since the former two types of leaders look less attractive and less warm than the latter two. Sports leaders could also be distinguished by the fact that they look less competent than business leaders, as well as more masculine and mature than the other types of leaders. However, as we explained above, one needs to be very careful not to misinterpret the implications of our findings or to draw invalid causal inferences from these correlational results. The results of our second study are provocative and interesting, but they will need to be replicated and investigated in more

detail. In the remainder of the paper, we consider the theoretical and practical implications of these findings.

Implications for Face-Based Leadership Inferences

First and foremost, these findings imply that, within several domains (business, military, and sports), individuals who achieve the highest positions of leadership share common facial features that distinguish them from leaders in other domains. This could occur for several reasons; for example, if these facial features are (actually) diagnostic of domain-specific leadership qualities or if these facial cues are ignored (by those who select leaders) but happen to correlate with other (non-facial) factors that do influence the leadership selection process. The most plausible explanation, in our view, is that leaders are being selected, at least partly, according to how they look. This is consistent with an extensive literature showing that appearances predict leadership success (e.g., Antonakis & Jacquart, 2013; Olivola & Todorov, 2010a). However, previous studies have mainly identified facial characteristics that are broadly associated with leadership success, even across domains. For example, both political and business leaders seem to benefit from having attractive or competent-looking faces (Antonakis & Dalgas, 2009; Berggren, Jordahl, & Poutvaara, 2010; Olivola & Todorov, 2010a; Pfann, Biddle, Hamermesh, & Bosman, 2000; Poutvaara et al., 2009; Rule & Ambady, 2008). In contrast, our results suggest that domain-specific facial stereotyping may also influence the leadership selection process, above-and-beyond other, more generic facial predictors of leadership success. That is, leaders may benefit not just from having competent-looking faces, but also from having facial features that "fit" a certain stereotype uniquely associated with

their particular domain. In fact, just having facial features that make one look like a good generic leader might not be sufficient to reach the most prestigious leadership positions in a domain; one may also need to possess facial features that stereotypically "fit" the leaders in that domain. The existence of domain-specific predictors of leadership success thus adds a new "twist" to our understanding of the role that face-based inferences play in leadership selection.

Interestingly, we also found that, whereas business, military, and sports leaders were distinguishable by their faces, political leaders (U.S. state Governors) were not. This suggests that leaders elected by the general population may have more variance in their faces and/or fewer identifying facial characteristics, compared to leaders (such as CEOs, U.S. Army Generals, and football coaches) who are selected by a relatively small group of like-minded peers. This could be due to the characteristics of the voting population: Compared to the three other groups of leaders in our study, U.S. state Governors (i) are elected by a larger, more diverse, and less uniformly-minded group of individuals, and (ii) they have less direct contact, and less in common, with the individuals who select them. Or it could be due to the nature of the leadership selection process: Gubernatorial elections are explicitly competitive and pit rival candidates representing opposing ideologies and views. To the extent that Republicans and Democrats associate different facial features with good political leadership (e.g., Olivola et al., 2012), we would expect elected leaders across the U.S. not to conform to a single facial stereotype. Each of these factors may, individually or in combination, limit the facial uniformity of (democratically elected) political leaders, relative to leaders in other domains. This connects nicely to previous work in the political domain showing that certain characteristics of the voting

population (Lenz & Lawson, 2011; Olivola et al., 2012) and the election process (e.g., how visible candidates are –see Olivola & Todorov, 2010a) seem to strengthen (or weaken) the predictive power of candidate facial features on election outcomes. Additional research comparing the facial characteristics of democratically elected and internally selected leaders could help shed light on the extent to which the organizational context (and, in particular, the leadership selection process) moderates the relationship between appearances and leadership success.

The striking dissociation we observed between subjective and objective performance has interesting implications as well. First, the fact that participants could not recognize when they were able to correctly infer leadership domain suggests that this judgment process might occur implicitly, at least to some extent. If people can (somehow) accurately distinguish the faces of business, military, and sports leaders without even realizing it, then perhaps the social, cognitive, and neural systems supporting these judgments operate below awareness. Second, these results imply that individuals (e.g., corporate board members or voters) who are most likely to trust, and therefore advocate, the use of facial cues to select leaders are no better at inferring leadership domain from faces than their (less confident) peers. In fact, to the extent that they rely too heavily on appearances at the cost of other, more diagnostic cues, these individuals may actually be less accurate than their peers (Olivola & Todorov, 2010b). As a result, the process of leadership selection risks being too heavily influenced by the poorly calibrated opinions of the most vocal believers in the diagnostic validity of faces.

Implications for Implicit Leadership Theories

The research reported in this paper may also shed some light on implicit leadership theories (or "ILTs"; see Lord, Foti, & Phillips, 1982; Lord, Foti, & DeVader, 1984; Offerman, Kennedy, & Wirtz, 1994) –the underlying assumptions, stereotypes, beliefs, and schemas that followers use to understand and evaluate leaders. Beyond predicting which individuals achieve prestigious leadership positions within particular domains, facial appearances likely influence the way these individuals are perceived by their followers once they assume leadership roles (e.g., Spisak, Dekker, Krüger, & van Vugt, 2012; Spisak, Homan, Grabo, & Van Vugt, 2012). The finding that leaders within certain domains share distinctive facial features could have several implications for ILTs. As we previously suggested, it might reflect a widely held (but perhaps erroneous) belief that certain facial features are indicative of good leadership within a particular domain. These face-based inferences about leadership quality or "fit" may, in turn, form an important component of ILTs and thus influence how followers perceive their leaders. From a leader's point of view, then, having facial features that make him look stereotypically "fit" to lead in his particular domain could be beneficial (above-and-beyond his actual fit), to the extent that it positively influences the way he is perceived by his followers (Brown, 2012). Being perceived as the "right" kind of leader, because of his face, can have implications for the leader's ability to persuade his followers to adopt goals and carry out tasks (Caldwell & O'Reilly, 1990; Pfeffer, 1981). For example, a leader might be seen as more "legitimate" and experience greater popularity if he possesses the facial features that people (including his followers) typically expect from leaders in his particular domain (Nye & Forsythe, 1991). By contrast, leaders who do not appear to "fit" with their leadership domain, based on their facial appearance, may arouse

skepticism and thus be required to compensate in other ways to achieve the same goals. Conversely, being repeatedly exposed to leaders with similar facial features may lead followers to associate leadership quality or "fit" with these superficial cues. The biasing influence of domain-specific facial stereotypes on ILTs could thus be self-reinforcing.

Unresolved Questions and Future Directions

The finding that human judges can correctly infer leadership domain from faces, seemingly without meta-cognitive awareness, naturally raises a number of interesting questions. One such question concerns the potential validity of these facial cues for judging domain-specific leadership quality (e.g., does looking like a stereotypical business leader actually indicate that one is particularly fit to lead a company?). We anticipate that a number of readers will interpret these findings to mean that people can accurately identify individuals best suited to lead within a particular domain, solely from their faces. However, our results do not imply that humans have an ability to accurately infer leadership quality or "fit" from facial cues. The fact that people can infer which domain a leader belongs to could just as well reflect the workings of a self-fulfilling bias, whereby individuals who possess certain facial features are *perceived* to be good leaders within a particular domain, and thus quickly promoted to leadership positions within that domain, regardless of their actual qualities (Antonakis, 2011). Indeed, several studies suggest that the relationship between facial appearance and leadership success is more likely to reflect biased perception than accurate inference (e.g., Graham, Harvey, & Puri, 2013; Lenz & Lawson, 2011; Olivola et al., 2012). Moreover, these same features could help leaders with the "right" looks advance their agendas within their organizations.

Attribution research shows that positive perceptions of leadership qualities can enhance a leader's ability to influence others (Calder 1977; Pfeffer, 1977). This, in turn, would give people the impression that they can trust their first impressions of leaders, thereby closing the self-fulfilling loop. In sum, erroneous folk theories regarding the diagnostic validity of faces combined with self-fulfilling pressures could spawn and nurture 'illusory correlations' (Chapman, 1967; Chapman & Chapman, 1969) between facial cues and leadership success (Antonakis, 2011). For these reasons, lay observers (and researchers) need to "differentiate between traits that *really* matter for leadership and those that *seem* to matter" (Antonakis, 2011, p. 272).

A second (and somewhat related) question concerns whether, and to what extent, having a face that stereotypically fits one's domain impacts other indicators of leadership success (beyond the attainment of a leadership position). In an initial effort to examine this question, we compared the vote shares obtained by U.S. Governors who were correctly identified by participants (and thus look like stereotypical politicians) with those who were less likely to be categorized as politicians. Among our Republican gubernatorial election winners, we found no relationship between looking like a politician (relative to other leadership categories) and vote shares obtained: $r(35) = -.02$, *ns*. In contrast, Democrat gubernatorial election winners who were more likely to be identified as politicians won by a smaller margin (of vote shares): $r(26) = -.45$, $p < .022$. Thus, looking like a politician does not seem to benefit political leaders (although, this may also depend on what *kind* of politician a candidate looks like –see Olivola et al., 2012). A similar analysis with our Fortune 500 CEOs (using revenue, profit, and salary data from 2012) produced mixed results: The likelihood of being correctly identified as a business

leader did not significantly predict their 1-year or 5-year earnings ($r(191) = .04$ and .11, respectively; both ps > .12), but it did predict the profit-to-revenue ratio of their companies in 2012 ($r(230) = .17$, $p < .01$). Clearly, more research will be needed to understand the potential benefits and/or costs of looking stereotypically like a leader from a particular domain.

A third set of questions concerns the relationship between facial stereotypes and leadership status (i.e., rank) within a domain. Our studies used the faces of individuals who had achieved some of the highest levels of leadership in their respective domains. How would our results look if we had instead used lower-ranked leaders? Do lower-ranked leaders have less distinctive faces? Would judges therefore perform worse at the task of identifying leadership domain from the faces of lower-ranked leaders? How would meta-accuracy differ (if at all) with lower-ranked leader faces as the stimuli? What about the facial stereotypes associated with each leadership domain –do these vary with leadership rank? We leave these questions for future research.

Conclusion

People seem to be surprisingly good at inferring leadership domain from facial cues, yet surprisingly bad at evaluating their ability to do so. These findings have clear implications for leadership perception and selection, yet more research is needed to better understand the potential impact that this kind of facial stereotyping might have on leaders and their followers.

References

Antonakis J. (2011). Predictors of leadership: The usual suspects and the suspect traits. In A. Bryman, D. Collinson, K. Grint, B. Jackson & M. Uhl-Bien (Eds.), *Sage Handbook of Leadership* (pp. 269-285). Thousand Oaks, CA: Sage.

Ames, D., Kammrath, L., Suppes, A., & Bolger, N. (2010). Not so fast: The (not-quite-complete) dissociation between accuracy and confidence in thin slice impressions. *Personality and Social Psychology Bulletin*, *36*, 264-277.

Antonakis, J., & Dalgas, O. (2009). Predicting elections: Child's play! *Science*, 323, 1183.

Antonakis, J., & Jacquart, P. (2013). The far side of leadership: Rather difficult to face. In M. C. Bligh, & R. E. Riggio (Eds.), *Exploring distance in leader-follower relationships: When near is far and far is near* (pp. 155-187). New York, NY: Routledge.

Avery, D. R., Tonidandel, S., Griffith, K. H., & Quinones, M. A. (2003). The impact of multiple measures of leader experience on leader effectiveness. *Journal of Business Research*, *56*, 673-679.

Bass, B. M., & Bass, R. (2008). *The Bass handbook of leadership: Theory, research, and managerial applications* (4th ed.). New York, NY: The Free Press.

Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences of the USA, 104*, 17948-17953.

Barnes, M. L., & Sternberg, R. J. (1989). Social intelligence and judgment policy of nonverbal cues. *Intelligence*, *13*, 263-287.

Berggren, N., Jordahl, H., & Poutvaara, P. (2010). The looks of a winner: Beauty and electoral success. *Journal of Public Economics*, *94*, 8-15.

Brown, D. J. (2012). In the minds of followers: Follower-centric approaches to leadership. In D. V. Day, & J. Antonakis, (Eds.), *The nature of leadership* (2nd ed.) (pp. 331-362). Thousand Oaks, CA: Sage.

Calder, B. J. (1977). An attribution theory of leadership. In Staw, B. M., and Salanclik, G. R. (Eds.), *New Directions in Organizational Behavior* (pp. 179-204). Chicago, Ill: St. Clair Press.

Caldwell, D. F. and O'Reilly, C. A. (1990). Measuring person-job fit with a profile-comparison process. *Journal of Applied Psychology*, *75*, 648-657.

Carpinella, C. M., & Johnson, K. L. (2013). Appearance-based politics: Sex-typed facial cues communicate political party affiliation. *Journal of Experimental Social Psychology*, *49*, 156-160.

Chapman, L. J. (1967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior*, *6*, 151-155.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, *74*, 271-280.

Cherulnik, P. D., Turns, L. C, & Wilderman, S. K. (1990). Physical appearance and leadership: Exploring the role of appearance-based attribution in leader emergence. *Joumal of Applied Social Psychology*, *20*, 1530-1539.

Davis, K. M., & Gardner, W. L. (2012). Charisma under crisis revisited: Presidential leadership, perceived leader effectiveness, and contextual influences. *The Leadership Quarterly*, *23*, 918-933.

Day, D. V., Sin, H., & Chen, T. T. (2004). Assessing the burdens of leadership: Effects of formal leadership roles on individual performance. *Personnel Psychology*, *57*, 573-605.

Efran, M. G., & Patterson, E. W. (1974). Voters vote beautiful: The effect of physical appearance on a national election. *Canadian Journal of Behavioral Science*, *6*, 352-356.

Eubanks, D. L., Antes, A. L., Friedrich, T. L., Caughron, J. J., Blackwell, L. V., Bedell-Avers, K. E., & Mumford, M. D. (2010). Criticism and outstanding leadership: An evaluation of leader reactions and critical outcomes. *The Leadership Quarterly*, *21*, 365-388.

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

Garland, D. J., & Barry, J. R. (1990). Personality and leader behaviors in collegiate football: A multidimensional approach to performance. *Journal of Research in Personality*, *24*, 355-370.

Giambatista, R. C. (2004). Jumping through hoops: A longitudinal study of leader life cycles in the NBA. *The Leadership Quarterly*, *15*, 607-624.

Graham, J. R., Harvey, C. R., & Puri, M. (2013). A corporate beauty context. Working Paper, Fuqua School of Business, Duke University.

Harms, P. D., Han, G., & Chen, H. (2012). Recognizing leadership at a distance: A study of leader effectiveness across cultures. *Journal of Leadership & Organizational Studies*, *19*, 164-172.

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

Hunter, S. T., Cushenbery, L., Thoroughgood, C., Johnson, J. E., & Ligon, G. S. (2011). First and ten leadership: A historiometric investigation of the CIP leadership model. *The Leadership Quarterly*, *22*, 70-91.

Jahoda, G. (1954). Political attitudes and judgments of other people. *Journal of Abnormal and Social Psychology*, *49*, 330-334.

Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychology*, *89*, 899-913.

Kalma, A. (1991). Hierarchisation and dominance assessment at first glance. *European Journal of Social Psychology*, *21*, 165-181.

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*, 574-589.

Ligon, G. S., Harris, D. J., & Hunter, S. T. (2012). Qualifying leader lives: What historiometric approaches can tell us. *The Leadership Quarterly*, *23*, 1104-1133.

Ligon, G. S., Hunter, S. T., & Mumford, M. D. (2008). Development of outstanding leadership: A life narrative approach. *Leadership Quarterly, 19,* 312-334.

Livingston, R. W., & Pearce, N. A. (2009). The teddy-bear effect: Does having a baby face benefit black chief executive officers? *Psychological Science*, *20*, 1229-1236.

Loehr, J., & O'Hara, R. B. (2013). Facial morphology predicts male fitness and rank but not survival in Second World War Finnish soldiers. *Biology Letters*, *9*, 20130049.

Lord, R. G., Foti, R. J., & De Vader, C. L. (1984). A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. *Organizational Behavior and Human Performance*, *34*, 343-378.

Lord, R. G., Foti, R. J., & Phillips, J. S. (1982). A theory of leadership categorization. In J. G. Hunt, U. Sekaran, & C. Schriesheim (Eds.), *Leadership: Beyond establishment views* (pp. 104-121). Carbondale, IL: Southern Illinois University Press.

Mazur, A., Mazur, J., & Keating, C. (1984). Military rank attainment of a West Point class: Effects of cadets' physical features. *American Journal of Sociology*, *90*, 125-150.

Mueller, U., & Mazur, A. (1996). Facial dominance of West Point cadets as a predictor of later military rank. *Social forces*, *74*, 823-850.

Mueller, U., & Mazur, A. (1997). Facial dominance in Homo sapiens as honest signaling of male quality. *Behavioral Ecology*, *8*, 569-579.

Mumford, M. D., Espejo, J., Hunter, S. T., Bedell, K. E., Eubanks, D. L., & Connelly, S. (2007). The sources of leader violence: A multi-level comparison of ideological and non-ideological leaders. *The Leadership Quarterly, 18,* 217-235.

Nye, J. L., & Forsythe, D. R. (1991). The effects of prototype-based biases on leadership appraisals: A test of leadership categorization theory. *Small Group Research*, *22*, 360-379.

Offermann, L. R., Kennedy, J. K., Jr., & Wirtz, P. W. (1994). Implicit leadership theories: Content, structure, and generalizability. *Leadership Quarterly*, *5*, 43-58.

Olivola, C. Y., Sussman, A. B., Tsetsos, K., Kang, O. E., & Todorov, A. (2012). Republicans prefer Republican-looking leaders: Political facial stereotypes predict candidate electoral success among right-leaning voters. *Social Psychological and Personality Science*, *3*, 605-613.

Olivola, C. Y., & Todorov, A. (2010a). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, *34*, 83-110.

Olivola, C. Y., & Todorov, A. (2010b). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology*, *46*, 315-324.

Pfann, G. A., Biddle, J. E., Hamermesh, D. S., & Bosman, C. M. (2000). Business success and businesses' beauty capital. *Economics Letters*, *67*, 201-207.

Pfeffer, J. (1977). The ambiguity of leadership. *Academy of Management Review*, *2*, 104-112.

Pfeffer, J. (1981). Management as symbolic action: The creation and maintenance of organizational paradigms. In L. L. Cummings & B. M. Staw (Eds.), *Research in organizational behavior* (Vol. 3). Greenwich, CT: JAI Press.

Poutvaara, P., Jordahl, H., & Berggren, N. (2009). Faces of politicians: Babyfacedness predicts inferred competence but not electoral success. *Journal of Experimental Social Psychology*, *45*, 1132-1135.

Reimers, S. (2009). A paycheck half-empty or half-full? Framing, fairness and progressive taxation. *Judgment and Decision-making*, *4*, 461-466.

Rule, N. O., & Ambady, N. (2008). The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological Science*, *19*, 109-111.

Schmid Mast, M., & Hall, J. A. (2004). Who is the boss and who is not? Accuracy of judging status. *Journal of Nonverbal Behavior*, *28*, 145-165.

Spisak, B. R., Dekker, P. H., Krüger, M., & van Vugt, M. (2012). Warriors and peacekeepers: Testing a biosocial implicit leadership hypothesis of intergroup relations using masculine and feminine faces. *PLoS-ONE*, *7*, e30399.

Spisak, B. R., Homan, A. C., Grabo, A., & Van Vugt, M. (2012). Facing the situation: Testing a biosocial contingency model of leadership in intergroup relations using masculine and feminine faces. *The Leadership Quarterly*, *23*, 273-280.

Stewart, L. H., Ajina, S., Getov, S., Bahrami, B., Todorov, A., & Rees, G. (2012). Unconscious evaluation of faces on social dimensions. *Journal of Experimental Psychology: General*, *141*, 715-727.

Todorov, A. (2012). The social perception of faces. In S. T. Fiske & C. N. Macrae (Eds.), *The SAGE Handbook of Social Cognition*. Thousand Oaks, CA: SAGE Publications.

Waldman, D. A., Ramirez, G. G., House, R. J., & Puranam, P. (2001). Does leadership matter? CEO leadership attributes and profitability under conditions of perceived environmental uncertainty. *Academy of Management Journal*, *44*, 134-143.

Wänke, M., Samochowiec, J., & Landwehr, J. (2012). Facial politics: Political judgment based on looks. In J. Forgas, K. Fiedler, & C. Sedikides (Eds.), *Social thinking and interpersonal behavior: Proceedings of the 14th Sydney symposium of social psychology* (pp. 143-160). New York, NY: Psychology Press.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after 100 ms exposure to a face. *Psychological Science*, *17*, 592-598.

Wong, E. M., Ormiston, M. E., & Haselhuhn, M. P. (2011). A face only an investor could love: CEOs' facial structure predicts their firms' financial performance. *Psychological Science*, *22*, 1478-1483.

Zebrowitz, L. A., & Collins, M. A. (1997). Accurate social perception at zero acquaintance: The affordances of a Gibsonian approach. *Personality and Social Psychology Review*, *1*, 203–222.

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass*, *2*, 1497–1517.

Notes

1) Here, we use the words "success", "successful", and "leadership success" to refer to the likelihood that a person is selected to a prestigious leadership position. To clarify, we are not referring to that person's leadership abilities and qualifications, nor to any successes he/she brings to their organization. The traits that make someone a popular candidate for a leadership position may well be different from those that make him/her a competent leader, once in that position. Our use of "success" (and its extensions) refers to the former (popularity), not the latter (competence).

2) There is also evidence that people can accurately infer a target's relative organizational status (Barnes & Sternberg, 1989; Schmid Mast & Hall, 2004) and behavioral indicators of dominance (Kalma, 1991), from facial photos.

3) Face-based judgments may also be partly spontaneous and perhaps difficult to control (Olivola & Todorov, 2010a; Stewart et al., 2012; Todorov, 2012). Therefore, even individuals who would rather avoid being influenced by appearances may be inadvertently affected, to some extent, by facial cues.

4) We discarded trials in which participants recognized one (or both) of the faces. In addition, a programming error led to a small proportion of photos (< 0.3%) being shown more than once to the same participants (these trials were also discarded). Finally, we discarded data from a few participants (n = 15) who saw the same photo presented three or more times (due to the programming error).

5) We correlated estimated accuracy (but not confidence) with the proportion of correct judgments across *all* trials within a block, including those with recognized or non-matched faces. This was done because participants were asked to estimate their performance over the entire set of trials in each block (i.e., we did not instruct them to ignore recognized or non-matched trials).

6) In line with previous studies (e.g., Ballew & Todorov, 2007; Olivola & Todorov, 2010a; Willis & Todorov, 2006), we instructed participants not to spend too much time forming their evaluations. The specific instructions they received were as follows: "We ask that you please rely on your ''gut feeling'' to form your impressions, without thinking too much about each face. There are no right or wrong answers; we are simply interested in your first, immediate impression of each person."

7) Along similar lines, Graham et al. (2013) find that CEOs have more competent-looking faces than non-CEOs. Our results show that CEOs are distinguishable, not only from non-leaders, but also from other kinds of leaders, by their facial competence.

Table 1. Study 1 experimental conditions

Condition	Block 1 Leadership category pairing		Block 2 Leadership category pairing	
1	*Business leaders* vs. Military leaders	B-M	*Political leaders* vs. Sports leaders	P-S
2	*Business leaders* vs. Military leaders	B-M	*Sports leaders* vs. Political leaders	S-P
3	*Business leaders* vs. Political leaders	B-P	*Military leaders* vs. Sports leaders	M-S
4	*Business leaders* vs. Political leaders	B-P	*Sports leaders* vs. Military leaders	S-M
5	*Business leaders* vs. Sports leaders	B-S	*Military leaders* vs. Political leaders	M-P
6	*Business leaders* vs. Sports leaders	B-S	*Political leaders* vs. Military leaders	P-M
7	*Military leaders* vs. Business leaders	M-B	*Political leaders* vs. Sports leaders	P-S
8	*Military leaders* vs. Business leaders	M-B	*Sports leaders* vs. Political leaders	S-P
9	*Military leaders* vs. Political leaders	M-P	*Business leaders* vs. Sports leaders	B-S
10	*Military leaders* vs. Political leaders	M-P	*Sports leaders* vs. Business leaders	S-B
11	*Military leaders* vs. Sports leaders	M-S	*Business leaders* vs. Political leaders	B-P
12	*Military leaders* vs. Sports leaders	M-S	*Political leaders* vs. Business leaders	P-B
13	*Political leaders* vs. Business leaders	P-B	*Military leaders* vs. Sports leaders	M-S
14	*Political leaders* vs. Business leaders	P-B	*Sports leaders* vs. Military leaders	S-M
15	*Political leaders* vs. Military leaders	P-M	*Business leaders* vs. Sports leaders	B-S
16	*Political leaders* vs. Military leaders	P-M	*Sports leaders* vs. Business leaders	S-B
17	*Political leaders* vs. Sports leaders	P-S	*Business leaders* vs. Military leaders	B-M
18	*Political leaders* vs. Sports leaders	P-S	*Military leaders* vs. Business leaders	M-B
19	*Sports leaders* vs. Business leaders	S-B	*Military leaders* vs. Political leaders	M-P
20	*Sports leaders* vs. Business leaders	S-B	*Political leaders* vs. Military leaders	P-M
21	*Sports leaders* vs. Military leaders	S-M	*Business leaders* vs. Political leaders	B-P
22	*Sports leaders* vs. Military leaders	S-M	*Political leaders* vs. Business leaders	P-B
23	*Sports leaders* vs. Political leaders	S-P	*Business leaders* vs. Military leaders	B-M
24	*Sports leaders* vs. Political leaders	S-P	*Military leaders* vs. Business leaders	M-B

Target categories are presented in italics (and on the left) within each category pair. The letters on the right indicate the abbreviated identification ‘code’ for each pair, with the first letter indicating the target category.

Table 2. The 15 dimensions of evaluation in Study 2 and their correlations

	Ambitious	Anxious	Attractive	Babyfaced	Charismatic	Competent	Confident
Ambitious							
Anxious	-.31**						
Attractive	.66***	-.53***					
Babyfaced	.26*	-.37***	.66***				
Charismatic	.51***	-.86***	.74***	.41***			
Competent	.73***	-.28*	.76***	.43***	.54***		
Confident	.51***	-.89***	.63***	.38***	.89***	.45***	
Conservative	.09	.64***	-.23*	-.16	-.50***	.17	-.52***
Disciplined	.20	.71***	-.12	-.18	-.48***	.30**	-.50***
Dominant	-.08	.79***	-.57***	-.53***	-.74***	-.28*	-.66***
Extraverted	.12	-.86***	.29**	.16	.76***	0	.80***
Likeable	.47***	-.84***	.76***	.52***	.93***	.61***	.84***
Masculine	.12	.25*	-.22*	-.67***	-.15	-.07	-.14
Threatening	-.30**	.82***	-.66***	-.54***	-.85***	-.49***	-.76***
Trustworthy	.42***	-.76***	.72***	.54***	.86***	.61***	.78***

	Conservative	Disciplined	Dominant	Extraverted	Likeable	Masculine	Threatening
Ambitious							
Anxious							
Attractive							
Babyfaced							
Charismatic							
Competent							
Confident							
Conservative							
Disciplined	.79***						
Dominant	.57***	.69***					
Extraverted	-.75***	-.80***	-.67***				
Likeable	-.43***	-.43***	-.81***	.66***			
Masculine	.08	.30**	.55***	-.14	-.29**		
Threatening	.41***	.50***	.87***	-.64***	-.92***	.39***	
Trustworthy	-.30**	-.34**	-.78***	.58***	.94***	-.29**	-.92***

Each coefficient represents the correlation between the average z-scores obtained by the (n = 80) leaders on a given pair of dimensions. Note: * $p < .05$; ** $p < .01$; *** $p < .001$.

Figure 1

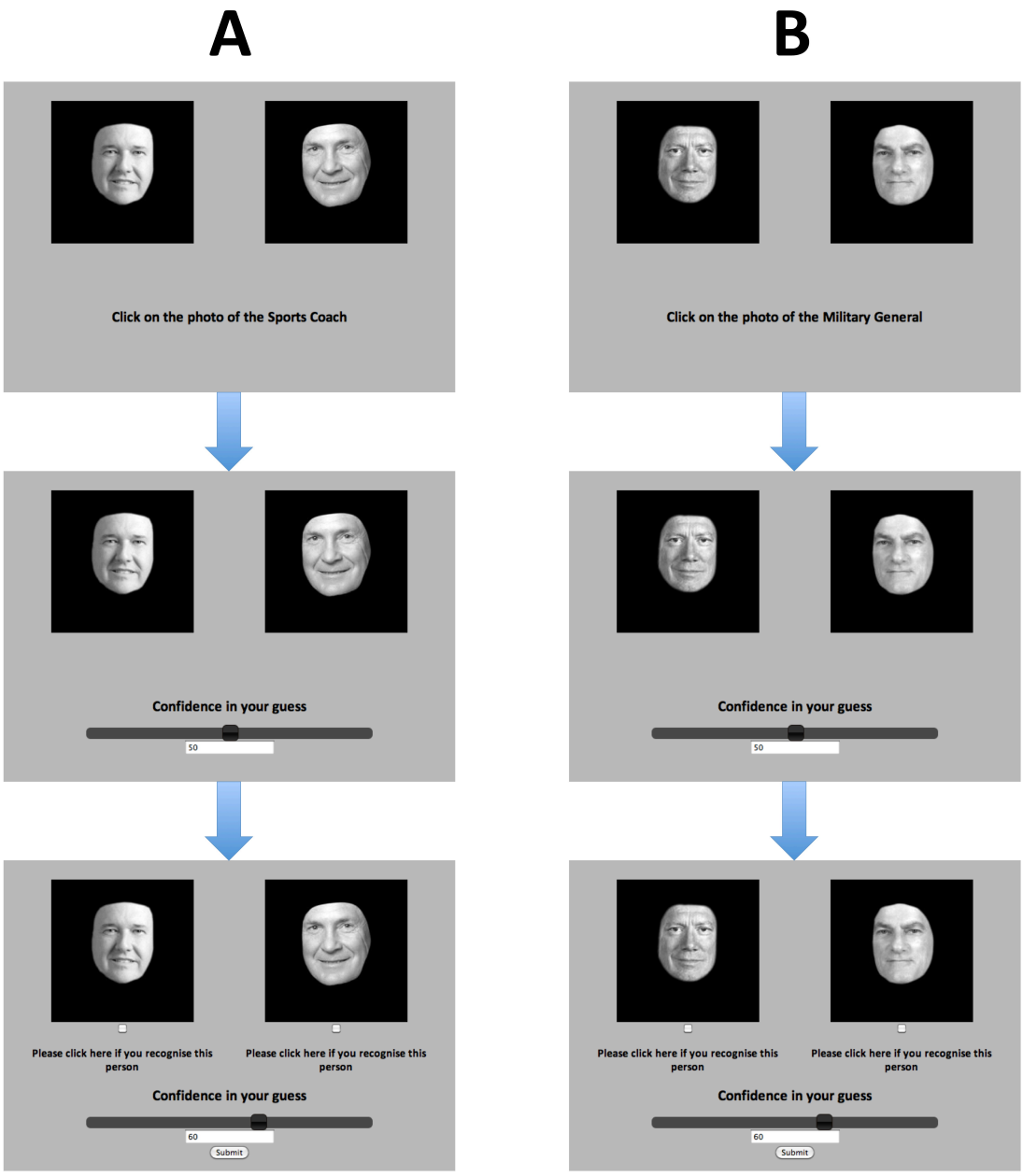

Figure 1.

Example screen shots from Study 1, showing the progression of a trial in two different conditions. The left column (A) shows a trial pitting sports leaders (American football

coaches) against business leaders (company CEOs), where the target category is sports leaders. The right column (B) shows a trial pitting military leaders (U.S. Army Generals) against political leaders (U.S. state Governors), where the target category is military leaders. First, participants indicated which person they thought belonged to the target category by clicking on his photo (top row). Next, they indicated how confident they were in their judgment (middle row). Finally, they reported whether they recognized one or both of the leaders (bottom row).

Figure 2

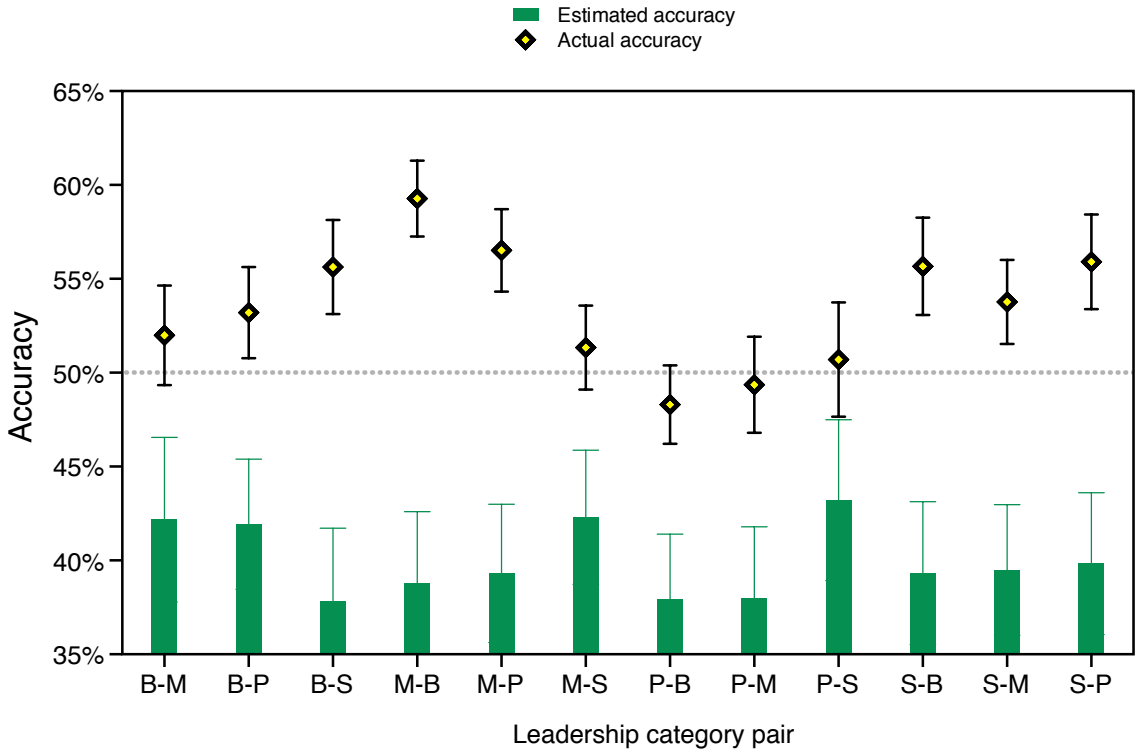

Figure 2.

Mean accuracy levels (in black) and estimated performance (in green) for each leadership category pairing (see Table 1). The dotted grey line represents chance-level performance, while error bars represent the 95% confidence intervals for the means. Therefore, mean accuracy levels are significantly different from chance if their error bars do not cross the dotted grey line.

Figure 3

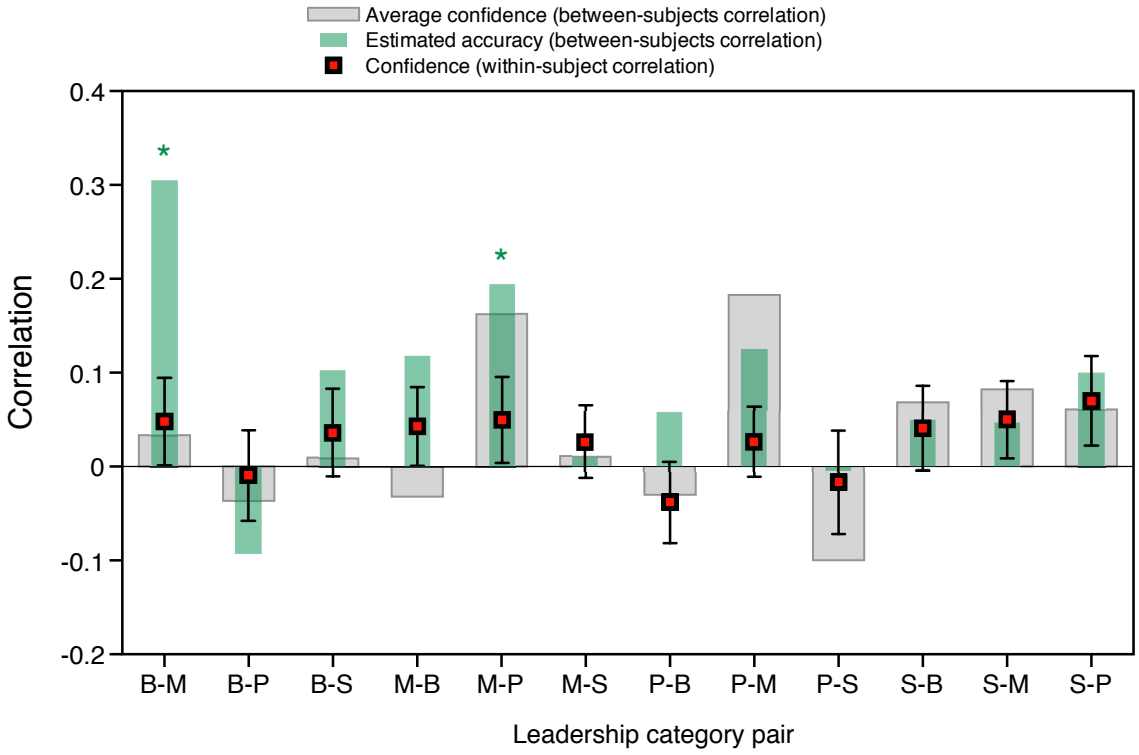

Figure 3.

Correlations between accuracy and estimated performance or accuracy and confidence, as a function of leadership category pairing (see Table 1). The thin green bars show the between-subject correlations between accuracy (for *all* trials – see Footnote 3) and estimated performance (with green asterisks indicating significant correlations at the $p < .05$ level). The wide grey bars show the between-subject correlations between accuracy and overall confidence (no correlations are significant at the $p < .05$ level). The red and black squares show the average within-subject (point-biserial) correlations between accuracy and confidence (with error bars representing the 95% confidence intervals).

Figure 4

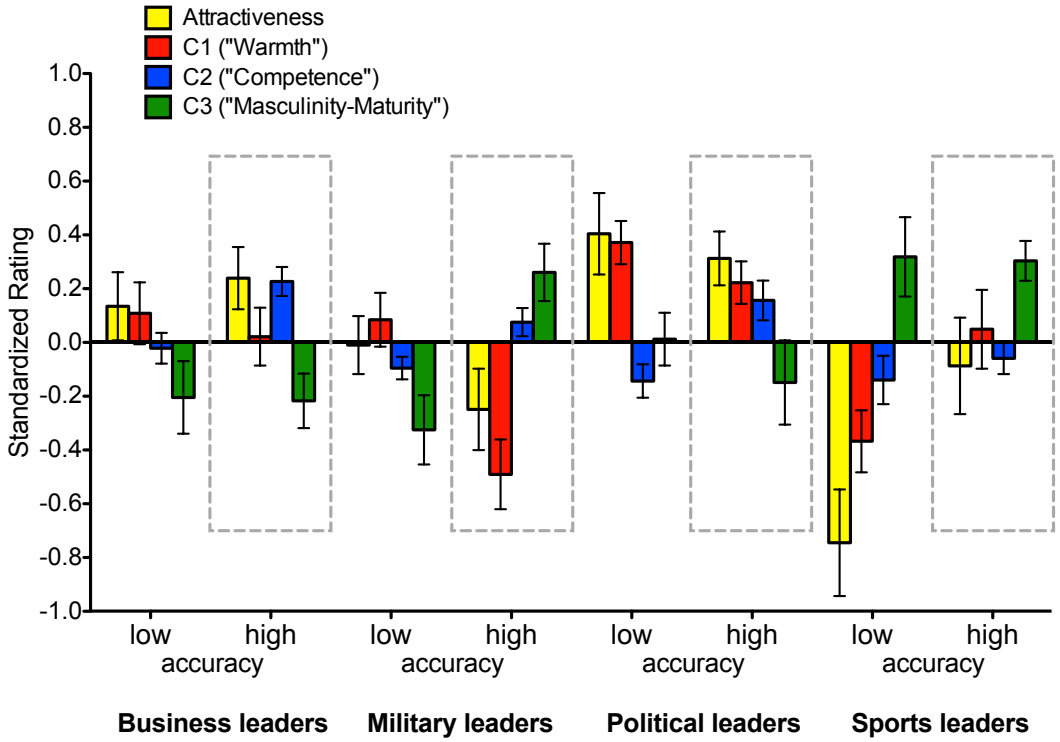

Figure 4.

Mean attractiveness and facial component scores for the 10 most accurately identified leaders (within grey dashed boxes) and the 10 least accurately identified leaders, as a function of leadership domain. Error bars represent standard error.

Figure 5

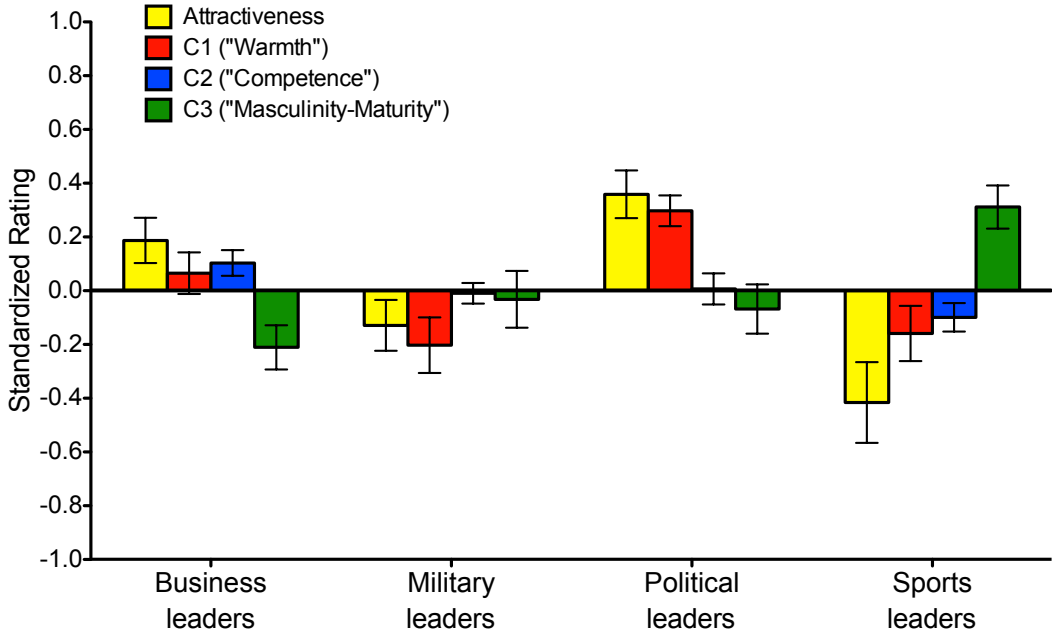

Figure 5.

Mean attractiveness and facial component scores as a function of leadership domain (combining the 10 most accurately identified and 10 least accurately identified leaders). Error bars represent standard error.

Appendix A – Principle Component Analysis (PCA) results (Study 2)

PCA 1 – All judgment variables entered

	Component 1 "Warmth"	Component 2 "Competence"	Component 3 "Masculinity-Maturity"
Ambitious	0.15	0.41*	0.29
Anxious	-0.31*	0.14	-0.12
Attractive	0.26	0.29	-0.04
Babyfaced	0.20	0.16	-0.50*
Charismatic	0.32*	0.04	0.18
Competent	0.17	0.47*	0.04
Confident	0.31*	0.00	0.22
Conservative	-0.19	0.38*	-0.13
Disciplined	-0.20	0.45*	0.03
Dominant	-0.30*	0.13	0.22
Extraverted	0.26	-0.30	0.21
Likeable	0.33*	0.08	0.03
Masculine	-0.12	0.05	0.68*
Threatening	-0.32*	-0.01	0.09
Trustworthy	0.31*	0.12	-0.02

Numbers represent the loadings of each judgment variable onto the three principle components. Asterisks indicate the variables that contributed to the calculation of a given evaluation dimension.

PCA 2 – Attractiveness excluded

	Component 1 "Warmth"	Component 2 "Competence"	Component 3 "Masculinity-Maturity"
Ambitious	0.14	0.44*	0.28
Anxious	-0.33*	0.10	-0.11
Attractive	.	.	.
Babyfaced	0.19	0.17	-0.50*
Charismatic	0.33*	0.08	0.17
Competent	0.16	0.52*	0.03
Confident	0.32*	0.04	0.21
Conservative	-0.21	0.39*	-0.14
Disciplined	-0.22	0.45*	0.02
Dominant	-0.31*	0.11	0.22
Extraverted	0.28	-0.28	0.21
Likeable	0.33*	0.13	0.02
Masculine	-0.13	0.04	0.68*
Threatening	-0.33*	-0.06	0.10
Trustworthy	0.32*	0.17	-0.03

Numbers represent the loadings of each judgment variable onto the three principle components. Asterisks indicate the variables that contributed to the calculation of a given evaluation dimension.

PCA 3 – Attractiveness, babyfacedness, and masculinity excluded

	Component 1 "Warmth"	Component 2 "Competence"
Ambitious	0.14	0.46*
Anxious	-0.34*	0.09
Attractive	.	.
Babyfaced	.	.
Charismatic	0.34*	0.10
Competent	0.16	0.53*
Confident	0.33*	0.06
Conservative	-0.22	0.39*
Disciplined	-0.23	0.45*
Dominant	-0.31*	0.11
Extraverted	0.30	-0.26
Likeable	0.34*	0.14
Masculine	.	.
Threatening	-0.33*	-0.07
Trustworthy	0.32*	0.18

Numbers represent the loadings of each judgment variable onto the three principle components. Asterisks indicate the variables that contributed to the calculation of a given evaluation dimension.