Research Article

Psychological Science
2014, Vol. 25(7) 1404–1417

DOI: 10.1177/0956797614532474
pss.sagepub.com

Misleading First Impressions: Different for Different Facial Images of the Same Person

Alexander Todorov[1,2] and Jenny M. Porter[3]
[1]Princeton University, [2]Radboud University Nijmegen, and [3]Columbia University

Abstract
Studies on first impressions from facial appearance have rapidly proliferated in the past decade. Almost all of these studies have relied on a single face image per target individual, and differences in impressions have been interpreted as originating in stable physiognomic differences between individuals. Here we show that images of the same individual can lead to different impressions, with within-individual image variance comparable to or exceeding between-individuals variance for a variety of social judgments (Experiment 1). We further show that preferences for images shift as a function of the context (e.g., selecting an image for online dating vs. a political campaign; Experiment 2), that preferences are predictably biased by the selection of the images (e.g., an image fitting a political campaign vs. a randomly selected image; Experiment 3), and that these biases are evident after extremely brief (40-ms) presentation of the images (Experiment 4). We discuss the implications of these findings for studies on the accuracy of first impressions.

Keywords
social perception, face perception

Received 5/7/13; Revision accepted 3/27/14

Once a relatively minor topic, social perception of faces has become the focus of a rapidly proliferating research literature (Todorov, Said, & Verosky, 2011; Zebrowitz, 2011). It is now well established that people make personality inferences from faces after minimal time exposure (Bar, Neta, & Linz, 2006; Todorov, Pakrashi, & Oosterhof, 2009; Willis & Todorov, 2006) and that these inferences predict important social outcomes (Flowe & Humphries, 2011; Graham, Harvey, & Puri, 2010; Olivola & Todorov, 2010a; Rezlescu, Duchaine, Olivola, & Chater, 2012; Rule & Ambady, 2008b).

The implicit assumption in studies of social perception of faces is that there is one-to-one mapping of image to personality. The image is treated as a faithful, unbiased representation of the individual's appearance that might provide diagnostic cues for his or her personality. In other words, it is assumed that different images of the same individual will result in the same personality judgments. However, this assumption is unwarranted. Burton and his colleagues have shown that minor variations in images of unfamiliar individuals substantially impair recognition of them (Burton & Jenkins, 2011). In a particularly telling demonstration, Jenkins, White, Van Montfort, and Burton (2011) presented British participants with 40 images of two Dutch celebrities, who were unknown to the participants. Participants were asked to sort the images into groups according to their identity. The modal response was nine identities. Not a single participant correctly categorized the images into two groups. Not surprisingly, this task was extremely easy for Dutch participants.

By definition, studies on first impressions are studies on judgments of strangers. The same considerations that apply to identity recognition apply to social judgments. That is, image variability may have large effects on such judgments. Jenkins et al. (2011) demonstrated this for perceptions of attractiveness. British participants rated the attractiveness of unfamiliar Dutch celebrities (20 celebrities × 20 images). Remarkably, for each pair of

Corresponding Author:
Alexander Todorov, Department of Psychology, Princeton University, Peretsman-Scully Hall, Princeton, NJ 08544
E-mail: atodorov@princeton.edu

celebrities, the one who was judged more attractive depended on the particular images whose ratings were compared. In fact, the within-person variability in judged attractiveness exceeded the between-person variability.

Similarly, different images of the same individual can result in different social attributions. In the experiments reported here, we tested whether minor, random variation in the images of the same individual could lead to different judgments. We used images produced for tests of face-recognition algorithms (Phillips, Moon, Rizvi, & Rauss, 1999): The individuals did not pose with any specific goal in mind (e.g., providing an image for online dating), and for all practical purposes, differences between images of the same individual could be treated as random.

In Experiment 1, for each of seven social judgments, the within-individual variance either was comparable to or exceeded the between-individuals variance. In other words, for almost any pair of individuals, the one rated higher on a given attribute depended on the images being considered. Experiment 2 shows that people have consistent preferences for specific images of a given person and that these preferences shift as a function of the decision context (e.g., political campaign vs. online dating). The judgments obtained in Experiment 1 predicted these preference shifts. Experiment 3 shows that people's judgments are predictably influenced by how the images they evaluate have been selected; their impressions are less positive if the images have been chosen randomly than if the images have been selected to best fit the decision context. Experiment 4 shows that such bias is evident even after extremely brief presentations of the images.

Experiment 1

Experiment 1 tested whether different images of the same individual result in different social attributions and perceptions of attractiveness. Participants judged the depicted individuals on attractiveness, competence, creativity, cunning, extraversion, meanness, trustworthiness, or intelligence. These attributes were selected because of their relevance to the context scenarios we used in Experiment 2.

Method

Participants. Eight hundred twenty people were recruited through Amazon Mechanical Turk and participated for payment. We did not use the data for 20 participants because test-retest correlations indicated that their judgments were unreliable (see Procedure). One hundred eight participants took more than one version of the survey; we used only the data from the first version they completed. The final sample consisted of 800 participants (mean age = 33.67 years, *SD* = 11.78 years; 416 females, 384 males).

Face stimuli. Pictures were acquired from CD 2 of the Face Recognition Technology (FERET) database of grayscale images (see Table S1 in the Supplemental Material available online for a list of the images and their names in the database; Phillips et al., 1999). This database was created to assist the development of automatic face-recognition algorithms and contains multiple head-shot images of each individual, with views ranging from frontal view to left and right profiles. In creating the database, 5 to 11 photographs were taken of each individual per session (Phillips et al., 1999; Phillips, Rauss, & Der, 1996). The database was developed to "measure the ability of the algorithms to handle large databases, changes in people's appearance over time, variations in illumination, scale, and pose, and changes in the background" (Phillips et al., 1996, p. 7).

Expressive behavior of the individuals was not manipulated, although there is variation in expressions. Specifically, if two consecutive frontal shots of an individual were taken in a single session, the individual was asked to present a different facial expression for the second shot than for the first, so that the images would not be identical. However, no direction was given as to the kind of expression desired or even that it should be an emotional expression. Thus, individuals generated spontaneous facial expressions with no constraints on what they should be other than different from one another. As a result, in the frontal images we used, the expressions do not vary systematically from the first to the second shot on a given day or across days. Furthermore, the facial expressions were not categorized or labeled by the individuals in the images, the photographers, or us in any way. For our purposes, the variation in expressive behavior is part of the natural variation in different images of the same individual.

We selected from the database 10 male and 10 female individuals who each had five different forward-facing photographs with adequate illumination of the face. The individuals were sampled from the first 100 individuals in the CD. Only 46 of those 100 had five or more frontal shots with good, consistent illumination across images. Within that sample of 46, the identities and specific images of each identity (if there were more than five frontal images) were selected at random, without taking facial expression into consideration. The resulting 100 images (20 individuals × 5 images) were cropped to remove most of the background. Figures S1 and S2 in the Supplemental Material present the full set of images. As the figures illustrate, expressions vary more for some individuals than for others. Note, however, that none of the images were selected by the individuals or the experimenters with a particular goal in mind.

To test whether the selected images of each individual could be identified as depicting that individual, we asked 10 participants (members of the Princeton community; mean age = 20.1 years; 6 females, 4 males) to classify the photographs. All 100 photographs were printed on 2- × 3-in. cards and randomly mixed together. Participants were seated at a table, and the pile of photo cards was placed in front of them. They were instructed to sort the cards into groups of pictures depicting the same person. Participants were not informed of the number of identities and were given unlimited time to complete the task. Although there were errors, they were not as pronounced as in the study by Jenkins et al. (2011), because we used images that had relatively restricted variability and could also be sorted by gender, age, and ethnicity. However, the pattern of errors was very similar to the pattern Jenkins et al. observed. The vast majority of errors (88%) were instances of classifying images of the same individual as images of different individuals (M = 3.6, SD = 3.81, range: 0–11). Participants rarely classified images of different individuals as images of the same individual (M = 0.5, SD = 0.97, range: 0–3). The median number of categories (individuals) was 21.5, and the mode was 20.

Procedure. A given participant judged only one image of a given individual, so that prior judgments of the same individual would not create biases. Each of the five different pictures of an individual was randomly assigned to a different image set. Thus, each of five sets contained exactly one picture of each person, and there were 20 pictures per set. Each participant was assigned to one of the five image sets and one of eight trait judgments (attractiveness, competence, creativity, cunning, extraversion, meanness, trustworthiness, or intelligence) in a 5 × 8 design. Participants judged how well each photo represented the intended trait (i.e., "How [trait] is this person?") on a 9-point scale ranging from 1 (*not at all* [trait]) to 9 (*extremely* [trait]). Participants were given unlimited time to respond to each face. However, they were instructed to go with their "gut instinct" and not spend too much time on any one face. Each picture was presented once in each of two blocks, with the order randomized within each block. Thus, participants rated each photo in the assigned image set twice. This allowed us to compute the test-retest reliability for each participant. Participants with zero or negative correlations between their test and retest judgments (n = 15) and participants with no variance in their judgments (n = 5) were replaced by new participants. The final sample consisted of 20 participants in each of the 40 cells.

Results

To the extent that subtle differences between images of the same individual are irrelevant to social judgments, judgments of images of the same individuals in different sets should be highly correlated. In contrast, if image differences matter, there should be wide fluctuations in the correlations, and this would suggest that the relative ordering of the individuals changes as their images change. The change in relative ordering should also be evident in comparisons of the variance of judgments of the same individual (across different images) with the variance of judgments of different individuals (with judgments averaged across the images of each individual). Specifically, the within-individual and between-individuals variances should be comparable.

For each of the 40 cells in our design (5 image sets × 8 judgments), the judgments were reliable (see Table S2 in the Supplemental Material). At the same time, as shown in Table 1, the range of correlations between judgments of different image sets (each depicting the same individuals) was large, from .03 to .87. The many low correlations indicate that the relative ordering of individuals changed as their images changed. These results cannot be attributed to low reliability of the judgments. For example, the reliability of extraversion judgments for all five image sets was equal to or exceeded .88 (see Table S2), yet the mean between-set correlation for the five image sets was only .20, and the correlation between sets was as low as .03.

Table 1 also shows the variance of judgments of the same individual (across different images) and the variance of judgments of different individuals (with judgments averaged across the images of each individual). The only judgment for which the between-individuals variance substantially exceeded the within-individual variance for both female and male faces was attractiveness. For all other judgments, with the exception of judgments of intelligence for male faces, the within-individual variance either was comparable to or exceeded the between-individuals variance. Across these latter seven judgments and across male and female individuals, the within-individual variance (M = 0.57, SD = 0.35) significantly exceeded the between-individuals variance (M = 0.35, SD = 0.12), $t(13) = 2.64$, $p < .02$.

Figure 1 shows the variations in judgments of attractiveness, extraversion, and trustworthiness for each individual (see Fig. S3 in the Supplemental Material for results for the remaining judgments). For each trait, the individuals are ordered according to the mean judgment across their five images. There was much less variation for judgments of attractiveness than for judgments of extraversion and trustworthiness. However, even in the case of attractiveness, there are many pairs of individuals for whom rankings reversed depending on the selection of images. It is also the case that the faces used were not perceived as attractive by the participants. The mean ratings for attractiveness were 3.33 and 3.90 for male and female faces, respectively (9-point scale). Thus, the range

Table 1. Correlations Between Judgments of Different Image Sets and Variance of Judgments of Different Individuals and of the Same Individual

	Correlation (*r*) between image sets[a]		Variance of judgments: female faces		Variance of judgments: male faces	
Judgment	Range	Mean	Between individuals	Within individuals	Between individuals	Within individuals
Attractive	.69–.87	.81	0.49	0.17	0.38	0.17
Competent	.18–.66	.48	0.33	0.31	0.46	0.39
Creative	.13–.67	.38	0.20	0.35	0.27	0.39
Cunning	.39–.78	.62	0.18	0.36	0.40	0.42
Extraverted	.03–.41	.20	0.55	1.44	0.35	0.99
Mean	.13–.65	.41	0.33	0.79	0.50	0.91
Smart	.35–.75	.57	0.26	0.27	0.52	0.28
Trustworthy	.44–.77	.62	0.22	0.63	0.34	0.42

[a]The different image sets depicted the same individuals.

of possible variation in the image judgments was constrained. Jenkins et al. (2011) found a much larger spread in attractiveness judgments of different images of the same individuals. However, they used images of celebrities, who are more likely to be perceived as attractive on average, and they used more images per individual without constraining differences between images. These factors would be expected to increase the within-individual variance of attractiveness judgments.

As shown in Figure 1, despite the small differences between images, for subjective judgments such as judgments of extraversion and trustworthiness, the within-individual variance is very large. In fact, the relative ranking of any pair of individuals on a given dimension can be reversed depending on the images selected. This is illustrated in Figure 2. Using the average ratings across the five images of each individual, we first identified the "most extraverted looking" and the "least extraverted looking" males and females, as well as the "most trustworthy looking" and the "least trustworthy looking." Nevertheless, we were able to find images of these individuals for which the "least extraverted" person was rated as more extraverted than the "most extraverted" person, and the "least trustworthy" person was rated as more trustworthy than the "most trustworthy" person.

As can be seen from Figure 2, this reversal of ratings was partly due to the expressive behavior of the individuals. Individuals who were smiling were rated as more extraverted and more trustworthy than those who were not. This pattern is consistent with both intuitions and computational models of impressions of extraversion and trustworthiness (Todorov, Dotsch, Porter, Oosterhof, & Falvello, 2013). However, we note three important points. First, our participants were judging a stable disposition (e.g., extraverted), not a momentary state (e.g., happy). If such dispositions are read from the face, they should not be easily swayed by momentary expressions. Second, expressive behavior is part of the natural variation of the images. As noted in the Method section, our images were sampled irrespective of any expressive behavior. Finally, even when we removed all images with smiling faces from analyses, the interpretation of our general findings did not change.

Indeed, it was easy to find reversals of ratings even among the images that did not show smiling faces (see Fig. 3). Thus, subtle differences in expressions lead to different (and inconsistent) judgments of stable dispositions. Not surprisingly, removing the smiling faces from the image sets reduced the within-individual variance ($M = 0.35$, $SD = 0.17$) and increased the between-individuals variance ($M = 0.45$, $SD = 0.16$) across the seven social judgments (see Table S3 in the Supplemental Material, which also shows results for the attractiveness judgments). As a result, the former did not exceed the latter. The difference between the two types of variance was not significant, $t(13) = 1.61$, $p = .13$. Thus, although open-smiling behavior contributed to the within-individual variance, it did not account for all of this variance. Moreover, such expressive behavior is part of the natural variation in the images.

Experiment 2

Experiment 2 tested whether different images of the same individual would be consistently favored in different contexts. Participants were provided with face images and asked to select the image that fit a particular context (e.g., online dating) best. To the extent that the particular image of an individual matters for such choices, the various images of a given individual should not be equally likely to be chosen. Moreover, to the extent that the context matters, preferences for images should shift as a function of context. Finally, these context-dependent preferences should be predicted by judgments of the images (Experiment 1). As we have argued elsewhere (Todorov et al., 2011), the decision context makes specific

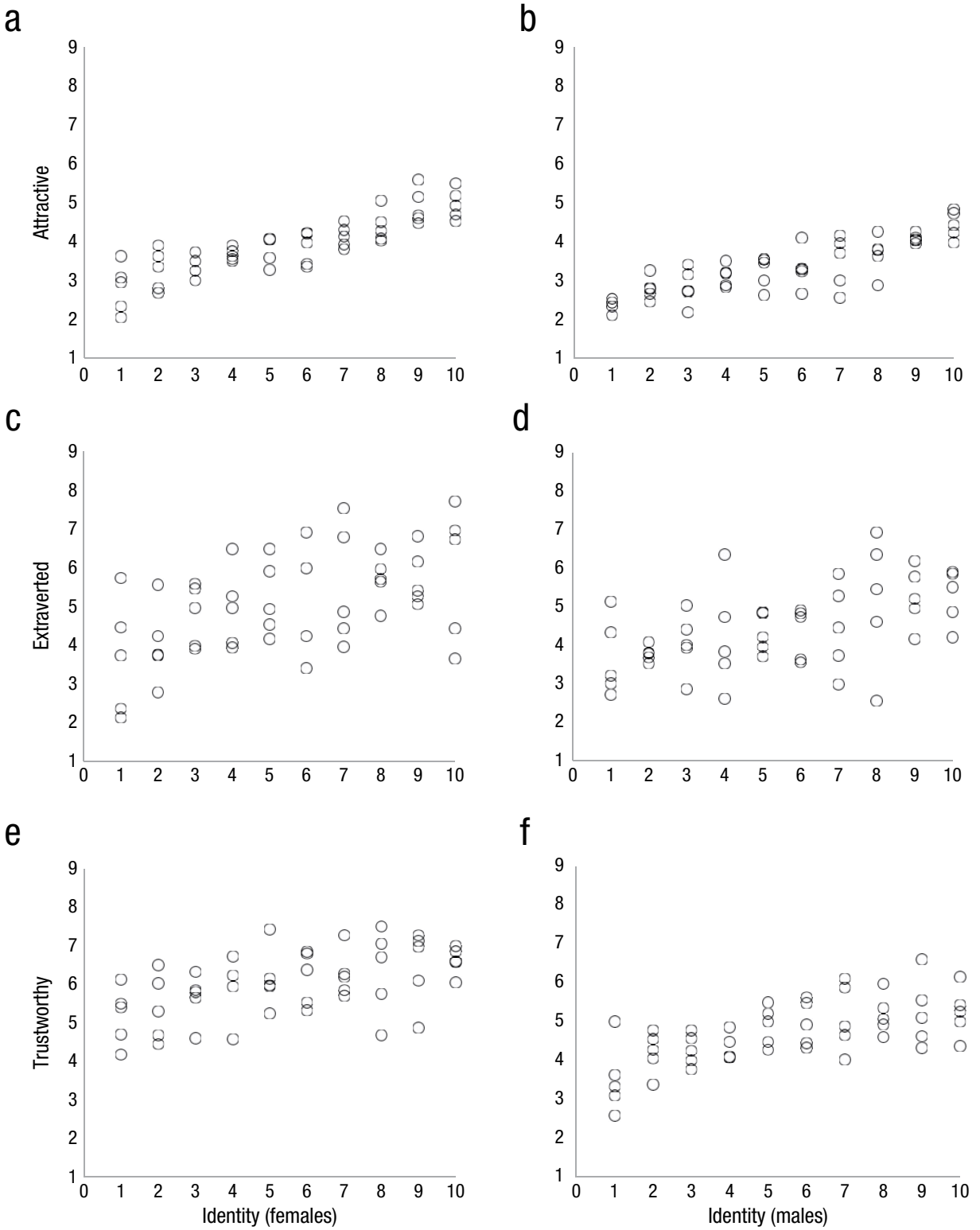

Fig. 1. Judgments of the 20 individuals in Experiment 1. Each circle represents the average rating of one of the five images of the indicated individual. Individuals are ordered on the *x*-axes according to their mean judgment across images. Each column of circles represents judgments of different images of the same individual. The graphs show perceived attractiveness of (a) female and (b) male faces, perceived extraversion of (c) female and (d) male faces, and perceived trustworthiness of (e) female and (f) male faces.

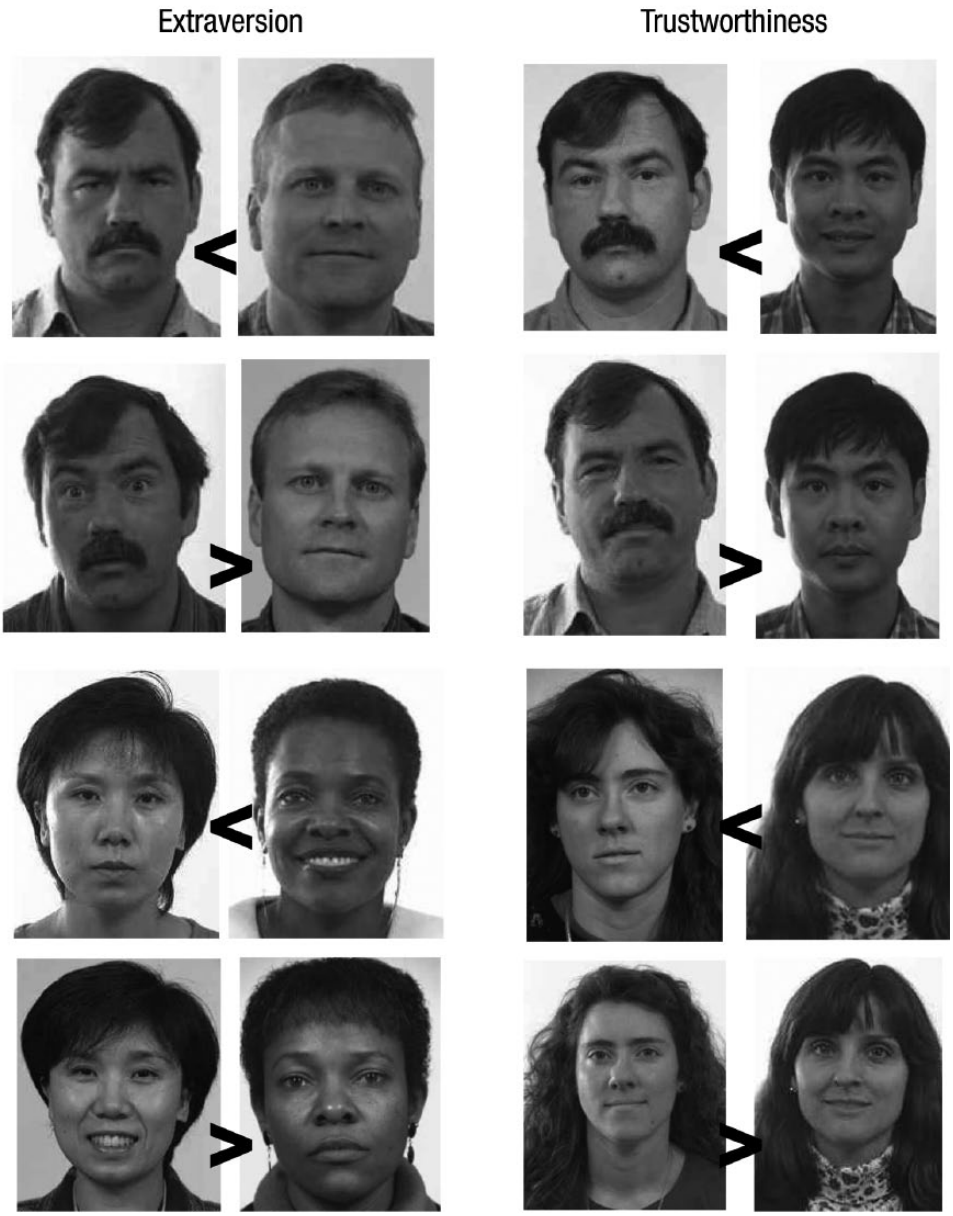

Fig. 2. Pairs of images illustrating reversals of judgments of extraversion and trustworthiness of the same individuals (Experiment 1). For each trait, the male and female individuals with the highest average ratings across five different images (right column) are paired with their counterparts who had the lowest average ratings (left column). For each pair, the top row shows images for which the individual with the highest average rating received a rating higher than the individual with the lowest average rating, and the bottom row shows images of these same individuals for which the relative ratings were reversed.

attributes important, and inferences of these attributes based on faces predict decisions.

Method

Participants. One hundred ten Princeton University students (53 females, 57 males; mean age = 19.58 years, *SD* = 1.22 years) participated for partial course credit or payment.

Procedure. We used the same face stimuli as in Experiment 1. Participants were randomly assigned to one of five scenarios (22 participants per scenario) before they viewed the photographs. The following five scenarios

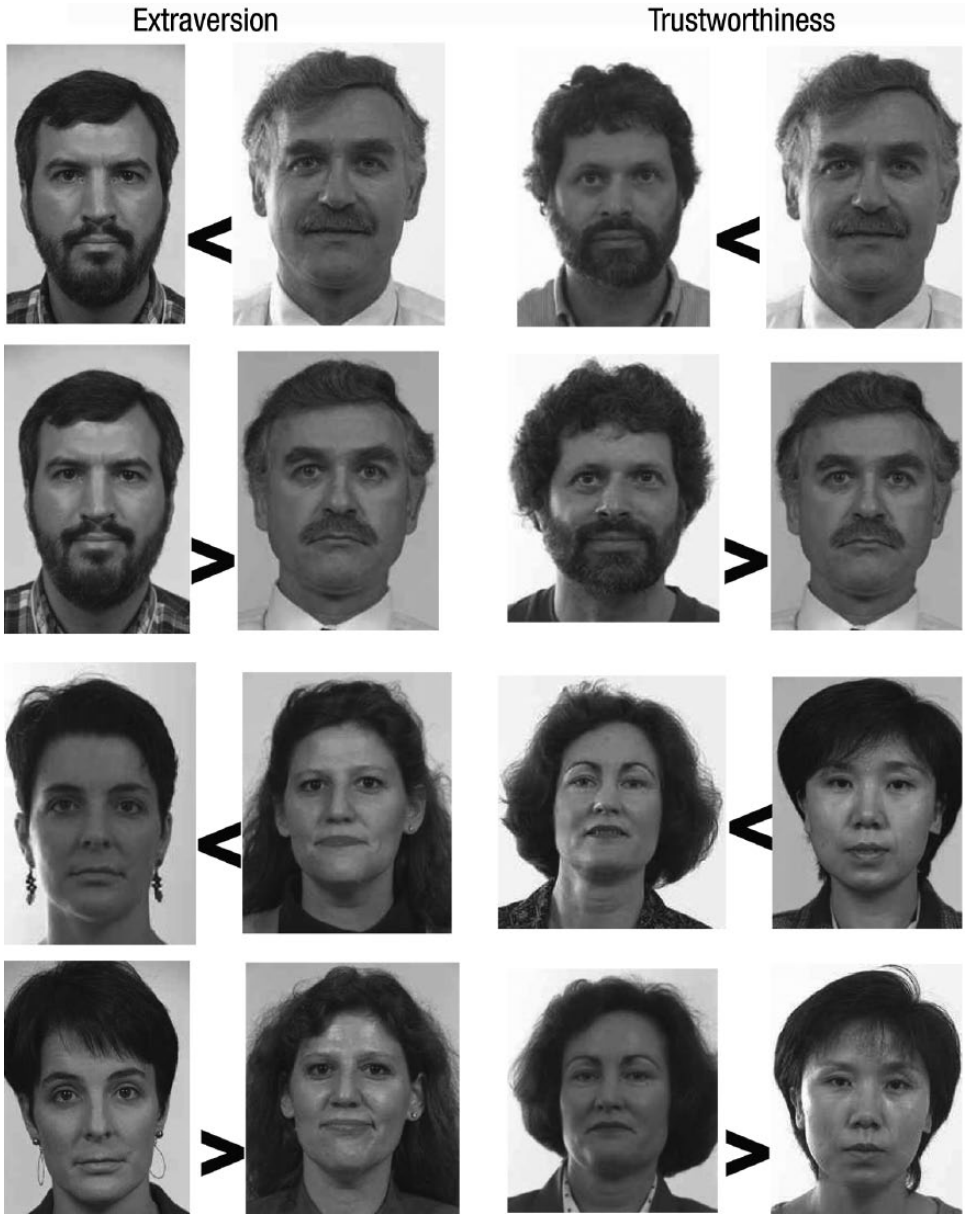

Fig. 3. Pairs of images illustrating reversals of judgments of extraversion and trustworthiness of the same individuals after faces showing open smiles were excluded (Experiment 1). For each pair, the top row shows images for which the individual on the right received a rating higher than the individual on the left, and the bottom row shows images of these same individuals for which the relative ratings were reversed.

were used: applying for a high-salary consulting position and selecting a photo for a resume, uploading a photo to an online-dating Web site, uploading a new Facebook photo, running for the office of town mayor and selecting a photo for campaign posters, and auditioning for the role of the villain in an upcoming film and selecting a photo for the application. On each trial, participants simultaneously viewed five photographs of a given individual, side by side, and were asked to imagine that they were the individual and to select the photo that best fit the scenario they had been given. All participants saw the same 20 individuals. The order in which the target individuals were presented and the positions of the photos on a given trial were randomized.

A 1,000-ms fixation cross preceded each trial. During each trial, participants viewed five different photos of the target face and read the following instructions: "Imagine you are this person. Keeping in mind the scenario, which photo would you choose?" Each photo had a number underneath (1–5). Participants selected their preferred

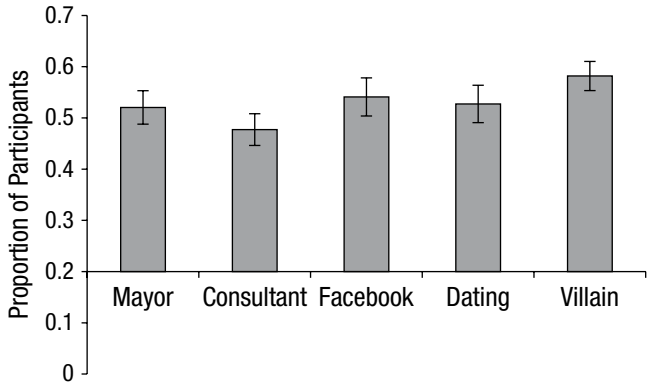

Fig. 4. Mean proportion of participants who made the dominant choice in each of the five scenarios in Experiment 2. The *x*-axis is at the chance level of .20. Error bars indicate ±1 *SEM*.

photo by pressing the number key that corresponded to the photo of their choice.

Results

If participants did not have any consistent preferences in their image choices, then the distribution of responses for each target individual and scenario would have been uniform—the expected distribution produced by random choices. This was clearly not the case. Out of 100 response distributions (20 target individuals × 5 scenarios), only 16 were not significantly different from a uniform distribution. (Note that significance was estimated from a relatively small sample of participants—n = 22 per specific choice—which reduced the probability of finding a significant effect.) The average chi-square value across individuals in the five scenarios ranged from 17.34, $p < .001$, to 25.84, $p < .001$. In summary, although the images of the target individuals were not taken with any specific social goal in mind, the natural variations in the images produced consistent preferences that favored some images over others. For each target individual and scenario, we identified the *dominant choice* (i.e., the image participants selected most often). As shown in Figure 4, within each scenario, the mean proportion of participants who made the dominant choice was significantly greater than would be expected by chance (.20), $t(19) > 8.94$, $p < .001$.

The second question of this study was whether different scenarios would lead to different choices of images. This was the case. We illustrate these results with the results for one of the target individuals (Fig. 5). For this man, the dominant choice for a campaign poster for a mayoral race was Image 1. Image 3, which was rarely chosen in this scenario, was the dominant choice for a Facebook photo. Image 2 was the dominant choice for both auditioning for the role of a villain and applying for a high-paying consulting position.

Thus, the dominant image choice changed as a function of the scenario. For example, the dominant choice in the online-dating scenario and the dominant choice in the villain scenario were different for all 20 target individuals. Across the 20 individuals, two to four different images dominated the choices in the five scenarios. For example, for the individual in Figure 5, three different images dominated the choices (Image 1 in the mayor scenario, Image 2 in the consultant and villain scenarios, and Image 3 in the Facebook scenario). The mean number of images that dominated the choices in the five scenarios[1] (M = 2.55, SD = 0.60) was significantly higher than 1, $t(19) = 11.46$, $p < .001$. Because it is possible that changes in the dominant choices were simply driven by the difference between the negative villain scenario and the four positive scenarios, we repeated these tests after excluding the villain scenario. Once again, the mean number of images that dominated the choices (M = 1.60, SD = 0.60) was significantly higher than 1, $t(19) = 4.49$, $p < .001$.

Finally, we used the judgments collected in Experiment 1 to predict the choices in Experiment 2. Specifically, for each of the 20 target individuals, we correlated the mean judgments of each image from Experiment 1 with the mean proportion of choices in Experiment 2. For each scenario and judgment, this generated 20 correlations that were Fisher-z-transformed, and the mean Fisher z score was tested against zero. Note that, if anything, the differences between the samples in Experiment 1 (Mechanical Turk participants) and Experiment 2 (Princeton students) should have contributed noise to this analysis and made it more difficult to detect significant effects. Nevertheless, 33 of the 40 correlations (8 judgments × 5 scenarios; see Table S4 in the Supplemental Material and Fig. 6) were significantly different from zero.

As shown in Figure 6, the pattern of correlations was highly systematic. First, higher positive judgments of images (e.g., trustworthy) corresponded to higher likelihood of selecting the images in positive scenarios (e.g., running for a mayor) and lower likelihood of selecting the images in the negative scenario (applying to be the villain in a movie). In contrast, higher negative judgments (mean and cunning) predicted the opposite pattern of choices. Second, specific judgments mattered more for some scenarios than for others. For example, the strongest predictors of choices in the consultant scenario were judgments of competence, intelligence, and trustworthiness, whereas the strongest predictors of choices in the online-dating scenario were judgments of trustworthiness, extraversion, and meanness (the latter being a negative predictor). Although attractiveness judgments significantly predicted the choices in all scenarios except for the villain scenario, they were not a

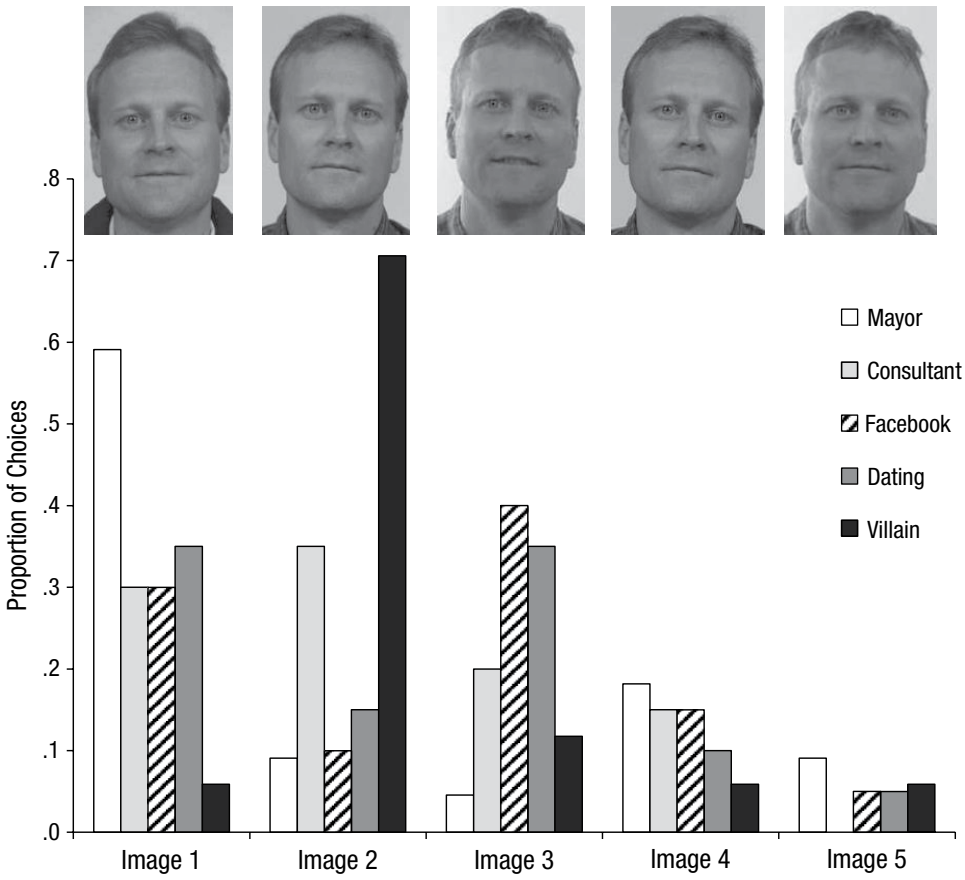

Fig. 5. Example of distributions of choices for one of the target individuals in Experiment 2. The graph shows the proportion of trials on which each image was selected in each of the five scenarios.

very strong predictor, most likely because the individuals depicted in the images were much older than the participants in both experiments and were not perceived as attractive.

Experiments 3a and 3b

Experiment 3 tested whether biasing the sample of images would affect decisions. Participants were presented with images identified as preferred in a specific context (consultant or mayor scenario from Experiment 2) and with randomly selected images, and were asked to make decisions based on the images (hiring or voting).

Method

Participants. Sixteen Princeton University students (7 females, 9 males; mean age = 19.81 years, *SD* = 1.42 years) participated in Experiment 3a for partial course credit. For Experiment 3b, data were collected from New Jersey residents (n = 22; 10 females, 12 males; mean age = 33.90 years, *SD* = 18.17 years); they were recruited in a local mall and were paid $2 for their participation.

Face stimuli. We used images of the same 20 individuals as in Experiment 1. In Experiment 3a, the 20 individuals were first randomly divided into two groups of 10. Then, for each individual, the image most frequently selected in Experiment 2 as the best photo in the consultant scenario was placed into a new group of the "best" images. An image of each individual was also randomly selected from among the five photos and placed into a new group of randomly selected images. The same procedure was followed for Experiment 3b except that the "best" images were the images most frequently selected as best in the mayor scenario in Experiment 2. Some of the randomly selected images were the same as the "best" images (one image for the consultant scenario and five images for the mayor scenario). Each participant saw one image of each target; half the images were from the "best" group and the

Fig. 6. Intensity color plot of the correlations between judgments from Experiment 1 and image choices in the five scenarios in Experiment 2. The judgments are ordered according to their weights on the first principal component resulting from a principal component analysis of the judgments (see Table S5 in the Supplemental Material). Asterisks indicate correlations significantly different from zero ($p < .05$).

other half were from the random group. These halves (from the groups of "best" and randomly selected images) were counterbalanced across participants.

Procedure. We constructed two decision scenarios based on the consultant and mayor scenarios in Experiment 2. In both scenarios, participants were asked to make a snapshot decision: In Experiment 3a, participants were asked to imagine that they were the person in charge of hiring at a corporation and to make a hiring decision based on their first impression. In Experiment 3b, participants were asked to imagine that they were voting in a local election and to make a voting decision based on their first impression. These scenarios were selected because participants' decisions were less likely to be influenced by differences in age, gender, and sexual orientation between participants and the target individuals than they would be in the dating and Facebook scenarios.

A 1,000-ms fixation cross preceded each trial. During each trial, participants saw a face image along with a scale from 1 (*not at all likely*) to 9 (*very likely*) and were asked to select the number corresponding to how likely they would be to hire the person pictured for the consulting position (Experiment 3a) or to vote for the person pictured for the office of mayor (Experiment 3b). The order of stimuli was randomized.

Results

We conducted analyses at both the level of participants and the level of the target individuals. The latter kind of analysis, in which responses are averaged across participants, is often the only kind reported in studies on social attributions based on facial appearance.

Participants in Experiment 3a were more likely to hire individuals whose photos were deliberately selected ($M = 5.41$, $SD = 0.79$) than individuals whose photos were randomly selected ($M = 4.87$, $SE = 0.80$), $t(15) = 3.42$, $p < .004$. This effect was also significant at the level of the target individuals, $t(19) = 2.91$, $p < .009$.

Participants in Experiment 3b were more likely to vote for individuals whose photos were deliberately selected ($M = 5.10$, $SD = 1.34$) than for individuals whose photos were randomly selected ($M = 4.76$, $SD = 1.48$), $t(21) = 2.01$, $p < .057$, although the difference was only marginally significant. This effect of selection was also marginally significant at the level of the target individuals, $t(19) = 1.98$, $p < .063$. The results of this experiment seem somewhat weaker than the results of Experiment 3a. However, as noted in the Method section, five of the images were the same in the "best" and random groups of images.

An analysis including both experiments, treating scenario as a between-subjects factor, showed a reliable

main effect of image selection for participants, $F(1, 36) = 13.47$, $p < .001$ (Fs < 1 for the main effect of scenario and the interaction of image selection with scenario). A similar analysis at the level of the target individuals also showed a main effect of image selection, $F(1, 19) = 10.15$, $p < .005$ (Fs < 1 for the main effect of scenario and the interaction).

Experiment 4

Experiment 4 tested whether participants form different judgments from different images of the same individual after extremely brief exposure to the images. Participants were presented with images that had been preferred either in a positive context (online dating) or in a negative context (villain movie) in Experiment 2 and were asked to rate the targets' likeability.

Method

Participants. Twenty-four Princeton University students and community members (17 females, 7 males; mean age = 22.46 years, $SD = 4.21$ years) participated for payment.

Face stimuli. The stimuli were the images most frequently selected for each target individual in Experiment 2 in the online-dating and villain scenarios. These scenarios were selected because the former is highly positive, the latter is highly negative, and the most frequent choices in the two scenarios were different for all 20 individuals. All images were resized to a height of 222 pixels, with variable width (to prevent stretching or distortion of the images).

Procedure. The experiment was run on 17-in. CRT monitors with a 75-Hz refresh rate. Participants viewed one image of each target individual, either the image most often chosen in the online-dating scenario or the image most often chosen in the villain scenario in Experiment 2. They rated how likeable each individual seemed, on a scale from 1 (*not at all likeable*) to 9 (*extremely likeable*). The image type (preferred for online dating or for the villain role) for each target individual was counterbalanced across participants. Participants were told that the faces would be presented for a very brief time and that they should indicate their first impression, or "gut" response.

The task consisted of one block of 20 trials. Each trial was preceded by a 1,500-ms blank screen, followed by a 500-ms fixation cross. During each trial, a target face was presented for 40 ms (three refresh cycles of the monitor) and immediately followed by a 200-ms mask consisting of gray-scale noise. Subsequently, the question "How likeable is this person?" appeared along with the rating scale.

Results

Despite the extremely brief presentation of the stimuli, participants liked the preferred images from the online-dating scenario ($M = 6.43$, $SD = 0.76$) much more than the preferred images from the villain scenario ($M = 4.55$, $SD = 1.03$), $t(19) = 10.54$, $p < .001$. Similarly, at the level of the target individuals, the online-dating images were much more likeable than the villain images, $t(19) = 10.28$, $p < .001$. In fact, this was the case for every target individual. At the level of the target individuals, we also tested whether differences in likeability judgments corresponded to differences in the valence of the images. Scores on the first principal component derived from the principal component analysis mentioned earlier (see Table S5) can be interpreted as valence evaluations, so to compute the valence difference between each pair of images, we computed the difference between their scores on the first principal component. The valence differences between images were strongly correlated with the differences in likeability judgments ($r = .68$, $p < .001$).

Discussion

Although we selected images that had similar lighting and face orientation and could be identified as depicting the same person, random variations in these images led to different personality impressions. Across social judgments, variability in impressions of the same individual was comparable to variability in impressions of different individuals (Experiment 1). Thus, the relative standing of two individuals on a given dimension depended on the images selected for comparison. Participants also had consistent preferences for specific images, and these preferences shifted as a function of the decision context (Experiment 2). For example, whereas some images were preferred in the context of a political campaign, others were preferred in the context of online dating. Not surprisingly, presenting participants with a biased selection of images consistently preferred (by other participants) in a given decision context resulted in biased judgments in a similar context (Experiment 3). For example, participants were more likely to vote for the person portrayed in a photo when that image had been the one preferred in a political-campaign context than when it had been chosen randomly. Finally, these biases were detectable even after extremely brief presentation of the images (Experiment 4).

Our findings dovetail with findings on person recognition and attractiveness judgments reported by Jenkins et al. (2011) even though we deliberately minimized the

differences among the images. We believe that if we had let the images vary on many additional parameters, such as lighting, face orientation, and head tilt, the differences in resulting person impressions would have been much larger than the ones we observed.

The fact that there is consensus in social judgments based on faces does not necessarily imply that these judgments are accurate (Hassin & Trope, 2000). The overgeneralization view of personality impressions, first proposed by Secord (1958) and then developed by Zebrowitz and her colleagues (McArthur & Baron, 1983; Zebrowitz, 2011), easily accommodates the current findings. Secord (1958) argued that people make inferences from momentary states of a person (e.g., a smile) and extend these inferences to attribute stable personality characteristics (e.g., friendly). For example, many studies have confirmed that people use similarity of unfamiliar faces to emotional expressions to make personality attributions (Montepare & Dobish, 2003; Neth & Martinez, 2009; Said, Sebe, & Todorov, 2009; Zebrowitz, Kikuchi, & Fellous, 2010). Although this research has exclusively compared images of different individuals, the same considerations apply to different images of the same individual. Different images capture different momentary states and, hence, can induce different personality inferences. The inferences may be accurate in the immediate context of the state, but would hardly generalize across different contexts.

Claims about the accuracy of personality inferences based on faces date back to ancient times and have been associated with the pseudoscience of physiognomy (Collins, 1999). Although the claims of the latter have been largely discredited, there has been renewed interest in questions about accuracy. Just in the past few years, there have been reports that people can accurately guess sexual orientation (Rule & Ambady, 2008a), political orientation (Rule & Ambady, 2010; Samochowiec, Wänke, & Fiedler, 2010; but see Olivola & Todorov, 2010b), and even criminal inclinations (Porter, England, Juodis, ten Brinke, & Wilson, 2008; Valla, Ceci, & Williams, 2011) from facial images alone.

Typically, participants in accuracy studies are presented with images of different individuals, known to differ on a specific attribute, and asked to guess the standing of the individuals on that attribute. Guessing that is better than chance is taken as evidence that the face conveys meaningful information about personality. However, the current findings suggest that if one cannot rule out selection biases in the images, one should be skeptical about accuracy claims. For example, many studies on the accuracy of judgments on sexual orientation have drawn their images from online-dating Web sites (Rule & Ambady, 2008a). It is quite likely that the Web-site users did not randomly select which images of themselves to post on these sites. Hence, it is possible that the presumed accuracy reflects biases in the selection of the images rather than honest or inherent signals of sexual orientation in the face. Findings that categorical judgments of sexual orientation can be made after extremely brief presentation of images (Rule & Ambady, 2008a) do not necessarily show that participants are sensitive to cues discriminating the categories. As suggested by the findings of Experiment 4, to the extent that there are biases in the selection of the images representing the two categories, the findings may simply indicate that participants are sensitive to the cues discriminating the two groups of images rather than the underlying categories per se.

Studies on the accuracy of judgments of criminal inclinations have involved comparisons of very different images: Mug shots of arrested and subsequently sentenced people have been compared with photographs of students on campus (Valla et al., 2011), and photographs of America's Most Wanted have been compared with photographs of Nobel Peace Prize winners (Porter et al., 2008). None of the control images in these studies were taken in the threatening and humiliating context of police arrest. This suggests that the "accurate" judgments may have to do more with the selection of images than with signals cuing criminal dispositions.

The current experiments certainly do not preclude the possibility that there may be accurate information in images. But, ultimately, finding accuracy is conditional on the selection of the images. Consider judgments of extraversion based on faces. Several studies have shown that such judgments correlate with actual extraversion (e.g., Borkenau & Liebler, 1992; Penton-Voak, Pound, Little, & Perrett, 2006). Yet extraversion was the judgment with the highest within-individual variance in our Experiment 1. Inspection of the images suggests that this variance resulted from an overreliance on smiling behavior (Fig. 2). We would expect to find correspondence between face judgments and extraversion to the extent that smiling is representative for the person across many situations. However, as demonstrated here, it would be easy to find images that lead to different conclusions. In short, it is difficult to make unambiguous statements about accuracy without knowledge of how images were produced and selected.

One can argue that these considerations do not apply to studies measuring invariant face characteristics. For example, studies have shown that the width-to-height ratio of male faces predicts aggressive behavior (Carré & McCormick, 2008; Carré, McCormick, & Mondloch, 2009). However, this measure is not truly invariant to image variation, as expressions, head tilt, and other variables can change this ratio. Further, several studies have failed to replicate the findings that this ratio predicts aggressive behavior (Deaner, Goetz, Shattuck, & Schnotala, 2012; Gómez-Valdés et al., 2012). Carré and his colleagues

recently reported that this ratio predicts aggressive behavior only in men with low social status (Goetz et al., 2013).

Conclusions

What we have shown here is something that people in the business of image manipulation have known for a long time. Yet most psychology research treats face images as veridical representations of individuals. This one-image-per-individual assumption has led to specific interpretations of findings on first impressions, namely, that differences in social attributions reflect interindividual face differences. However, the face is not a still image frozen in time but rather a constantly shifting stream of expressions that convey different mental states. To the extent that these mental states lead to different personality inferences, single snapshots captured in still images of faces may be a poor source of accurate personality inferences.

Author Contributions

A. Todorov designed the experiments. J. M. Porter collected the data. A. Todorov and J. M. Porter analyzed the data. A. Todorov wrote the first draft of the manuscript, and both authors edited it.

Acknowledgments

We thank Virginia Falvello for her help with Experiment 1.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

Open Practices

All data and materials can be accessed at http://tlab.princeton.edu/publications/todorovandporterdataandstimuli/. The materials are also available in the Supplemental Material. The complete Open Practices Disclosure for this article can be found at http://pss.sagepub.com/content/by/supplemental-data.

Note

1. Ties were coded conservatively against our hypothesis. For example, if Image 1 dominated the choices in the consultant scenario, and Images 1 and 2 were tied in the Facebook scenario, we considered there to be just one dominant choice in the two scenarios.

References

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*, 269–278.

Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology*, *62*, 645–657.

Burton, A. M., & Jenkins, R. (2011). Unfamiliar face perception. In A. Calder, J. V. Haxby, M. Johnson, & G. Rhodes (Eds.), *Handbook of face perception* (pp. 287–306). New York, NY: Oxford University Press.

Carré, J. M., & McCormick, C. M. (2008). In your face: Facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proceedings of the Royal Society B: Biological Sciences*, *275*, 2651–2656.

Carré, J. M., McCormick, C. M., & Mondloch, C. J. (2009). Facial structure is a reliable cue of aggressive behavior. *Psychological Science*, *20*, 1194–1198.

Collins, A. F. (1999). The enduring appeal of physiognomy: Physical appearance as a sign of temperament, character, and intelligence. *History of Psychology*, *2*, 251–276.

Deaner, R. O., Goetz, S. M. M., Shattuck, K., & Schnotala, T. (2012). Body weight, not facial width-to-height ratio, predicts aggression in pro hockey players. *Journal of Research in Personality*, *46*, 235–238.

Flowe, H. D., & Humphries, J. E. (2011). An examination of criminal face bias in a random sample of police lineups. *Applied Cognitive Psychology*, *25*, 265–273.

Goetz, S. M. M., Shattuck, K. S., Miller, R. M., Campbell, J. A., Lozoya, E., Weisfeld, G. E., & Carré, J. M. (2013). Social status moderates the relationship between facial structure and aggression. *Psychological Science*, *24*, 2329–2334.

Gómez-Valdés, J., Hünemeier, T., Quinto-Sánchez, M., Paschetta, C., Azevedo, S., González, M. F., . . . González-José, R. (2012). Lack of support for the association between facial shape and aggression: A reappraisal based on a worldwide population genetics perspective. *PLoS ONE*, *8*(1), Article e52317. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0052317

Graham, J. R., Harvey, C. R., & Puri, M. (2010). *A corporate beauty contest* (National Bureau of Economic Research Working Paper 15906). Retrieved from http://www.nber.org/papers/w15906.pdf?new_window=1

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

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*, 313–323.

McArthur, L. A., & Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological Review*, *90*, 215–238.

Montepare, J. M., & Dobish, H. (2003). The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of Nonverbal Behavior*, *27*, 237–254.

Neth, D., & Martinez, A. M. (2009). Emotion perception in emotionless face images suggests a norm-based representation. *Journal of Vision*, *9*(1), Article 5. Retrieved from http://www.journalofvision.org/content/9/1/5.full

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.

Penton-Voak, I. S., Pound, N., Little, A. C., & Perrett, D. I. (2006). Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social Cognition*, *24*, 607–640.

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (1999). *The FERET evaluation methodology for face-recognition algorithms* (National Institute for Standards and Technology Internal Report No. 6264). Retrieved from http://www.itl.nist.gov/iad/humanid/feret/doc/feret_methodology_nist_ir_6264.pdf

Phillips, P. J., Rauss, P. J., & Der, S. Z. (1996). *FERET (Face Recognition Technology): Recognition algorithm development and test results* (ARL-TR-995). Adelphi, MD: Army Research Laboratory. Retrieved from http://www.itl.nist.gov/iad/humanid/feret/doc/army_feret3.pdf

Porter, S., England, L., Juodis, M., ten Brinke, L., & Wilson, K. (2008). Is the face the window to the soul?: Investigation of the accuracy of intuitive judgments of the trustworthiness of human faces. *Canadian Journal of Behavioural Science*, *40*, 171–177.

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLoS ONE*, *7*(3), Article e34293. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0034293

Rule, N. O., & Ambady, N. (2008a). Brief exposures: Male sexual orientation is accurately perceived at 50 ms. *Journal of Experimental Social Psychology*, *44*, 1100–1105.

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

Rule, N. O., & Ambady, N. (2010). Democrats and Republicans can be differentiated from their faces. *PLoS ONE*, *5*, Article e8733. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0008733

Said, C., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion*, *9*, 260–264.

Samochowiec, J., Wänke, M., & Fiedler, K. (2010). Political ideology at face value. *Social Psychological & Personality Science*, *1*, 206–213.

Secord, P. F. (1958). Facial features and inference processes in interpersonal perception. In R. Tagiuri & L. Petrullo (Eds.), *Person perception and interpersonal behavior* (pp. 301–315). Stanford, CA: Stanford University Press.

Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion*, *13*, 724–738.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, *27*, 813–833.

Todorov, A., Said, C. P., & Verosky, S. C. (2011). Personality impressions from facial appearance. In A. Calder, J. V. Haxby, M. Johnson, & G. Rhodes (Eds.), *Handbook of face perception* (pp. 631–652). New York, NY: Oxford University Press.

Valla, J. M., Ceci, S. J., & Williams, W. M. (2011). The accuracy of inferences about criminality based on facial appearance. *Journal of Social, Evolutionary, and Cultural Psychology*, *5*, 66–91.

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

Zebrowitz, L. A. (2011). Ecological and social approaches to face perception. In A. Calder, J. V. Haxby, M. Johnson, & G. Rhodes (Eds.), *Handbook of face perception* (pp. 31–50). New York, NY: Oxford University Press.

Zebrowitz, L. A., Kikuchi, M., & Fellous, J. M. (2010). Facial resemblance to emotions: Group differences, impression effects, and race stereotypes. *Journal of Personality and Social Psychology*, *98*, 175–189.