Fooled by First Impressions? Reexamining the Diagnostic Value of Appearance-Based Inferences

Christopher Y. Olivola, Alexander Todorov

Fooled by First Impressions? Reexamining the Diagnostic Value of Appearance-Based Inferences

Christopher Y. Olivola, Alexander Todorov

PII: S0022-1031(09)00308-4
DOI: 10.1016/j.jesp.2009.12.002
Reference: YJESP 2390

To appear in: *Journal of Experimental Social Psychology*

Received Date: 21 August 2009
Revised Date: 19 November 2009

Please cite this article as: C.Y. Olivola, A. Todorov, Fooled by First Impressions? Reexamining the Diagnostic Value of Appearance-Based Inferences, *Journal of Experimental Social Psychology* (2009), doi: 10.1016/j.jesp.2009.12.002

Running Head: Fooled by First Impressions

Fooled by First Impressions?
Reexamining the Diagnostic Value of Appearance-Based Inferences

Christopher Y. Olivola[1] & Alexander Todorov[2]

1) University College London
2) Princeton University

Address Correspondence to:

Chris Olivola
Cognitive, Perceptual and Brain Sciences
University College London
26 Bedford Way
London WC1H 0AP
United Kingdom
(+44) (0)20-7679-7570 (phone)
(+44) (0)20-7436-4276 (fax)
c.olivola@ucl.ac.uk

CITATION COUNT: 40

KEYWORDS: person perception; social cognition; judgment accuracy; nonverbal behavior; first impressions; spontaneous trait inferences; web-based research

Abstract

We often form opinions about the characteristics of others from single, static samples of their appearance --the very first thing we see when, or even before, we meet them. These inferences occur spontaneously, rapidly, and can impact decisions in a variety of important domains. A crucial question, then, is whether appearance-based inferences are accurate. Using a naturalistic data set of more than 1 million appearance-based judgments obtained from a popular website (Study 1) and data from an online experiment involving over a thousand participants (Study 2), we evaluate the ability of human judges to infer the characteristics of others from their appearances. We find that judges are generally less accurate at predicting characteristics than they would be if they ignored appearance cues and instead only relied on their knowledge of characteristic base-rate frequencies. The findings suggest that appearances are overweighed in judgments and can have detrimental effects on accuracy. We conclude that future research should (i) identify the specific visual cues that people use when they draw inferences from appearances, (ii) determine which of these cues promote or hinder accurate social judgments, and (iii) examine how inference goals and contexts moderate the use and diagnostic validity of these cues.

"Beware, as long as you live, of judging people by appearances."

--*The Cockerel, the Cat, and the Young Mouse* (Jean de La Fontaine, 1668/1974)

Despite the old adage warning us not to "judge a book by its cover," we often form opinions about the characteristics of others from single, static samples of their appearance --the very first thing we see when, or even before, we meet them (Hassin & Trope, 2000; Todorov, Said, Engell, & Oosterhof, 2008; Zebrowitz, 1996). These inferences occur spontaneously and rapidly (Ballew & Todorov, 2007; Bar, Neta, & Linz, 2006; Rule & Ambady, 2008a; Todorov, Pakrashi, & Oosterhof, in press; Willis & Todorov, 2006). Furthermore, recent evidence suggests that these impressions impact the decisions that people make in a variety of important domains, including mate choice (Olivola et al., 2009), politics (for reviews of this literature see: Hall, Goren, Chaiken, & Todorov, 2009; Olivola & Todorov, in press), business/finance (Gorn, Jiang, & Johar, 2008; Naylor, 2007; Pope & Sydnor, 2008; Ravina, 2008; Rule & Ambady, 2008b), law/forensic-science (Blair, Judd, & Chapleau, 2004; Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006; Zarkadi, Wade, & Stewart, in press; Zebrowitz & McDonald, 1991), and the military (Mueller & Mazur, 1996).

A crucial question, then, is whether appearance-based inferences are valid forms of social judgment. That is, can we use appearances to determine a target-person's characteristics, or are we being fooled by first impressions? The answer to this question has serious and wide-ranging implications. The widespread use of visual media and the growing popularity of the Internet mean that appearances are increasingly the first cues

we receive about another person (e.g., through posted photos), often long before we meet them.

While previous studies have examined the diagnostic validity of appearances, the resulting evidence has been mixed (Ambady, Hallahan, & Conner, 1999; Hassin & Trope, 2000; Rule & Ambady, 2008a; Zebrowitz & Collins, 1997; Zebrowitz & Montepare, 2008). Furthermore, in many of these studies, the distributions of target characteristics were manipulated to be equiprobable, and thus did not reflect actual category membership frequencies in the real world. This feature, in particular, may have led to premature and overly optimistic conclusions regarding the diagnostic value of appearances in everyday social judgments – a point that we return to in the discussion.

The goal of this paper is to critically explore the validity of appearance-based judgments by examining what happens when one can draw on both appearances *and* category frequency information to infer something about another person. A competent judge with access to both pieces of information should weigh each cue in proportion to its validity and thus perform (on average) as well as or better than she would if she only had access to one of them. If, however, we tend to allocate too much weight to appearances, then the availability of photos and other static social-visual cues may actually hinder our ability to form accurate social judgments about characteristics with highly predominant categories. In this case, reliance on appearances could be detrimental to judgment accuracy, even when appearance-based inferences are "accurate" in the sense that they exceed chance. To explore this possibility, we conducted two studies in which we measured people's judgment accuracy-levels as they tried to guess others' characteristics using photos of the targets and information about the underlying category frequencies. In

Study 1, we compare performance across a variety of characteristics that differ naturally in terms of their category-frequency distributions. In Study 2, we focus on a single characteristic but experimentally vary its category-frequencies to see how this impacts performance.

Study 1

In Study 1, we used a large naturalistic data set containing over 1 million appearance-based judgments, produced over the course of a year (the site was launched in May 2005 and the data were collected in May 2006). These were obtained from a popular website (*www.whatsmyimage.com*), which allows users to predict specific facts about each other from their photos. In fact, the stated purpose of the site is to allow people to discover what kind of impression they convey through their appearance. This is made clear in the site's name ("What's My Image?") and its mission statement, which reads as follows:

"It is often said that first impressions are lasting impressions. Do you ever wonder what first impression strangers draw from you? What assumptions do people make about you before they learn the truth?

"What's My Image?" is a novel website to help you find the answer. Here, you can upload photos of yourself and then ask complete strangers to make guesses about the details of your private life. These are facts that no one could possibly determine from your photo, so the stranger's guess is entirely based on your image."

Website users interested in having others predict their characteristics from their appearance simply posted photos of themselves, chose which characteristics (from a list) they wanted others to guess, and reported which categories they fell into for these characteristics. Others could then view these photos and guess the category that each target fell into. Judges could choose to view pictures of men, women, or both. On each "trial", a judge was presented with one photo of a target and a randomly selected characteristic to predict (see Figure 1). After each prediction, a new target and characteristic were randomly selected. Judges also received immediate feedback concerning the accuracy of each prediction and the distribution of others' guesses, giving them an opportunity to learn the overall frequencies (i.e., base-rates) of categories. Finally, judges earned points for correct guesses, with the highest scorers featured in a "hall of fame" scoreboard on the site --an additional incentive to maximize accuracy.

Figure 1 about here

Data-set & Methods

The initial sample consisted of 901 targets, who received a combined total of 1,005,406 guesses about their characteristics from their posted photos. We focused our analysis on perceptually ambiguous characteristics defined by clear categories[1]. This led us to select 11 characteristics (see Table 1): 10 binary (yes/no) variables and one variable (sexual orientation) with three categories (heterosexual, homosexual, and bisexual).

Many targets in this dataset received a large number of guesses per characteristic, with some receiving more than 2,000 for a single characteristic. We measured the mean guessing accuracy for each target-characteristic combination. Photos receiving fewer than 10 guesses were discarded from the analysis. Although the majority of targets only posted a single photo, 30% posted several (between two and five) pictures of themselves. In these latter cases, guessing accuracy was averaged across a target's various photos before being included in the analyses. Finally, a few targets posting multiple photos provided inconsistent data (e.g., a person who reported drinking in one photo but not in another). Data on all such targets were discarded, but *only* for the specific questions to which they provided inconsistent answers. Our analysis in Study 1 is thus at the level of targets and characteristics, not at the level of judges or individual photos.

Table 1 reports, for each characteristic, the exact question that was posed on the website, the number of target users that were selected for our analysis (based on the method of selection described above), the most frequent category that targets fell into, and the proportion of selected target users who were male, in college, and/or working full-time, at the time the data were collected.

Table 1 about here

We also measured prior beliefs by surveying 98 undergraduate students (43% male; age range: 18-23, M = 19.65, SD = 1.23) about the category they believed

American adults most frequently fell into for each characteristic.

We used these data to calculate three statistics for each characteristic:

- Website performance: the mean accuracy, across targets, of judges on the website.

- Dominant base-rate: the proportion of targets falling into the *most frequent* category (based on the initial sample of all target users). This benchmark corresponds to the accuracy-level that would be achieved by judges who knew the base-rates and guessed the most frequent category on every trial.

- Survey performance: the mean accuracy that our survey respondents would achieve on the website by consistently guessing the categories they believed to be most frequent of American adults. To the extent that survey-respondents' and website-users' beliefs overlap, this benchmark tells us the accuracy-level that the latter group could achieve by ignoring photos and feedback, and relying solely on their prior beliefs. For an *n*-category characteristic with X_i% of targets falling into category *i* and Y_i% of survey respondents believing *i* to be most frequent, this accuracy-level would equal:

$$\sum_{i=1}^{n}\left(\frac{X_i}{100}\bullet\frac{Y_i}{100}\right)$$

To the extent that judges properly weigh the diagnostic value of appearances, prior

beliefs, and feedback, they should outperform survey respondents, and possibly the dominant base-rate.

Figure 2 about here

Results

We found that male and female targets reported very similar characteristics, the only two major differences being that male users were much more likely to report having been in a fist fight and/or owning a gun. However, accuracy-levels were, for the most part, comparable for both target genders, regardless of the characteristic being judged (the largest difference was only 7%). Given these similarities, we decided to collapse the data across target gender. The statistics for each characteristic are presented in Figure 2. Across all characteristics, the mean accuracy of website judges significantly exceeded chance-levels (chance levels were: 33% for sexual orientation and 50% for all other characteristics). However, to conclude from this result that appearances are valid cues for social inference assumes that judges had no information about the distributions of characteristics. Judges likely had some prior knowledge about the underlying base-rates. Indeed, the correlation between the dominant base-rates and mean accuracy-levels was .73, suggesting that judges were sensitive to underlying frequencies. While the optimal prediction strategy integrates base-rate information with any additional evidence obtained from target-photos (using Bayes' rule), this may be too difficult for most people (Hastie

& Dawes, 2001). A simpler strategy is to consistently guess the most frequent category for each characteristic. Yet, for all but one characteristic, observed performance fell significantly below the accuracy expected from this dominant base-rate strategy. Erroneous prior beliefs is an unlikely explanation[2] since our survey respondents could have outperformed website judges on 7 out of 11 characteristics by consistently relying on their subjective prior-beliefs about category-frequencies, despite the fact that website judges had two additional sources of information: photos and feedback. We would therefore expect website judges to do *at least as well* as survey respondents, not worse.

Another possible explanation for this failure is the well-documented tendency for people to "probability match" (Erev & Barron, 2005; Estes, 1972) -- to guess each category in proportion to its experienced frequency-- rather than consistently guess the most frequent category. Consider, for example, a characteristic such as sexual orientation, which has a dominant base-rate of 90%. For a judge who knows only the category base-rates (and nothing else), the optimal strategy would be to consistently guess "heterosexual" on every trial (i.e., 100% of the time), which would guarantee an accuracy-level close to 90%. However, if this person were probability matching, then they would instead guess "heterosexual" 90% of time, and guess "bisexual" or "homosexual" on the remaining 10% of trials, which would produce an accuracy level below 90% (closer to 82%). Probability matching behavior of this sort might therefore appear to explain why performance falls below the optimum. However, there are at least two reasons to be skeptical of this account. First, in contrast to experimental designs that typically yield probability matching behavior, the website judges were asked to make predictions about multiple variables across trials rather than to consistently guess the

outcome of a single discrete variable. The random ordering of characteristics across trials ensured that judges were not consistently guessing or receiving feedback about the same variable, which likely hindered probability matching. Second, the data do not support a probability matching account. For 5 out of 11 characteristics, we found that a probability matching strategy would have significantly outperformed website judgments. Furthermore, for only 3 out of 11 characteristics did the accuracy-level that would be achieved by probability matching actually fall within the 95% confidence interval observed for website judgments. Thus probability matching cannot fully account for the results we observe.

Discussion

Study 1 provides compelling evidence that appearances hinder the use of other information. For most characteristics, we found that our survey respondents (who did not see the photos) could have outperformed the website judges, even though this latter group actually had more information to rely on. Only when the dominant base-rate approached equiprobability (50%) did website judges perform above the level achievable by survey respondents. In fact, in the case of the "College degree" characteristic, which got closest to uniform priors, website performance even exceeded the dominant base-rate.

Curious readers may wonder why "College degree" was the only characteristic for which the performance of website users significantly exceeded the dominant base-rate. This might reflect previous evidence that people can (under some circumstances) accurately judge intelligence from facial cues (Zebrowitz, Hall, Murphy, & Rhodes, 2002; Zebrowitz & Rhodes, 2004). Alternatively, this could simply be a Type I error, resulting from the combined fact that (i) the dominant base-rate for "College degree" was

very low (52%), (ii) the dominant base-rate is just below the 95% confidence interval for accuracy, and (iii) since we examined 11 characteristics, the likelihood of obtaining a false significant difference is relatively high. But the simplest explanation is that, when determining whether targets had college degrees, judges were aided by the fact that some targets may have been noticeably too young to have already graduated from college (e.g., noticeably younger than 20 years old). Thus, cues about age (which is largely inferable from appearance) could have allowed judges to perform better on this task than on others.

Collecting data from a website provides a number of advantages that are difficult to achieve in a standard laboratory setting. One obvious example is the large sample size, but another important one is ecological validity. In most studies that have examined the accuracy of appearance-based inferences, the participants knew they were taking part in research. This knowledge, and the expectations that accompany it, could have affected how they approached the judgment task, including their motivation to be accurate and their beliefs about important experimental parameters (such as the base-rates). By contrast, it is safe to assume that the website judges in Study 1 were mostly responding in an environment natural to them, and without the direct scrutiny of experimenters.

Despite its advantages, Study 1 also has a number of limitations. First, we had no control over the base-rates associated with each characteristic, as these reflected the natural frequencies of category occurrence in the population (to the extent that website users were mostly honest and representative of the general population). A characteristic's dominant base-rate might be correlated with how difficult it is to accurately infer from appearances, in which case the two variables would be confounded. Second, the photos that users posted varied considerably in terms of their content and quality. Although, in

this paper, we are interested in appearances more generally, it would be interesting to see whether our results replicate with more standardized photos, for which there are fewer appearance cues distinguishing targets. Third, judges chose how many trials they wanted to complete before quitting and they could also return at a later time to participate in more trials. This could have amplified differences in how tired, motivated, and attentive they were on any given trial, thereby adding variance to their performance. Finally, Study 1 does not allow us to rigorously tease-apart the contributions of appearances, feedback, and prior beliefs on judgment accuracy. The feedback information, which included both the distribution of others' guesses and the correct answer, is particularly difficult to analyze and may have produced some (small) amount of probability matching behavior among judges.

In order to deal with these limitations, Study 2 used an experimental design that allowed us to systematically and independently vary both the base-rates and knowledge of these base-rates. In addition, the stimuli used and the number of trials that participants completed were standardized to reduce additional sources of variance. To simplify the task, participants only had to guess one characteristic (political affiliation) rather than alternating randomly between multiple characteristics. Finally, no feedback was provided after each trial, and only two types of cues were available to judges: appearances and (for half the participants) the base-rates of the two categories that targets could fall into.

Study 2

For Study 2, we programmed and conducted an online experiment in the guise of a "Political Guessing Game", in which participants tried to guess the political affiliation

of U.S. politicians solely from their facial photos. Previous studies conducted in the U.K. have shown that people agree in their guesses of political affiliation from photos (Bull & Hawkes, 1982; Bull, Jenkins, & Stevens, 1983) and that these guesses could exceed chance (Jahoda, 1954). As an incentive to participate, judges were informed that they would receive feedback, at the end of the experiment, about how well they performed.

Methods

Participants were mainly recruited through a link in an article (see Olivola & Todorov, 2009) on *Scientific American's* popular "Mind Matters" website. At the end of the article (which discussed the impact of physical appearances on political success), readers were invited to test their ability to guess the political affiliation (Democrat or Republican) of various American political candidates from these politicians' facial photos. Specifically, the two final paragraphs of the article read as follows:

An important, and as yet unanswered, question concerns the accuracy of judgments based on facial appearances: Are competent-looking politicians actually more competent than their not-so-competent-looking rivals? Or, more broadly, can you tell something about a political candidate solely from his or her appearance? Play our Political Guessing Game to find out!

In this game you will be presented with photos of political candidates and asked to guess their political affiliation. Once you finish the game, you can find out how well you were able to distinguish Republicans and Democrats by their appearance. In addition, your participation will help answer important questions about the human ability to draw information from the faces of politicians.

The article provided a link to the "Political Guessing Game" website, where the experiment was presented to participants.

Participants

Through this link, we collected 1,018 sets of responses. After excluding data from participants who reported being younger than 18 or having participated already, our final sample consisted of 1,005 participants (30% female; Age: *Range* = 18-87, *M* = 36.40, *SD* = 14.32). The modal participant was a Democrat with U.S. citizenship, currently living in the United States, who had voted in at least one American election, and who did not recognize any of the politicians in the study.

Stimuli

The politicians presented in this study were all candidates from the 2002 and 2004 House of Representatives elections. Candidates for the House of Representatives were chosen because they receive less media exposure and are thus less recognizable than Senate, gubernatorial, or presidential candidates, yet their photos are still publicly accessible in most cases. In addition, there are many candidates for the House of Representatives (there are 435 Representatives, and elections occur every two years), which provides us with many stimuli. The candidate photos were headshots drawn from a set of standardized stimuli that we had previously used in other studies (Olivola & Todorov, in press; Todorov, Mandisodza, Goren, & Hall, 2005; see the latter reference for details on the procedures involved in obtaining, selecting, and standardizing these photos). Highly recognizable candidates (e.g., Jesse Jackson, Jr.; Bobby Jindal; Ron Paul) were excluded, as were those whose photos were of low quality or who were turned away

from the camera in their photo. These photographs were transformed to black-and-white bitmap files and standardized in size (width = 3.2 cm, height = 4.5 cm). Any conspicuous background (e.g., the Capital or a U.S. flag) was removed and replaced with gray background. The photos were then separated into four pools according to politician gender and political party (only Democrats and Republicans were selected). Our final stimuli set consisted of 784 political candidate photos: 98 female Democrats, 61 female Republicans, 296 male Democrats, and 329 male Republicans.

Procedure

We independently varied three factors, all between participants. First, half the participants were shown only female political candidates, while the other half were shown only male candidates. Second, the proportion of candidates who were Democrats was varied between 10% and 90%, in increments of 10%. Finally, half the participants were informed of this proportion, while the other half were simply told that the proportion of Democrats could be equal to, smaller than, or larger than, the proportion of Republicans. The experiment was thus a 2 (politician gender) by 9 (proportion of Democrats) by 2 (base-rate information) between-subjects design. Participants were assigned to one of the 36 resulting conditions in alternating order.

The data were collected in May and June of 2009, over a 5-week period (although, over 90% of our data were collected in the first ten days and half were collected in the first two days). The experiment was conducted entirely through the Internet and consisted of 60 trials. On each trial, a photo of a different political candidate was presented in the center of the screen, and participants had to guess whether the person was a Democrat or a Republican by clicking on the appropriate label below the

picture (the two labels presented below each photo were simply the words "Democrat" and "Republican"). The next trial was presented immediately after the participant responded. The 60 candidates that a given participant saw were randomly drawn from the appropriate pools of photos, according to the gender and base-rate condition that this person was assigned to. For example, for a participant assigned to guess the political affiliation of female candidates under a 70% Democrat base-rate, 42 candidates were randomly drawn from the entire pool of female Democrats and 18 were randomly drawn from the entire pool of female Republicans. The order in which these candidates were presented was then randomized. In addition, the location of the response labels was randomly determined for each participant at the beginning of the experiment and held constant across the 60 trials. In other words, approximately half of our participants saw the label "Republican" located on the right and the label "Democrat" located on the left, while this ordering was reversed for the other half.

Before they could begin, participants had to provide informed consent by reading and checking a statement at the bottom of the introductory webpage, which explained the nature and purpose of the study. Participants were then presented with the instructions. First, a webpage with general instructions asked participants to read the instructions and questions carefully, and to refrain from talking to anyone else during the study. Then a second webpage explained the specific features of the study. Participants were told that they would be shown 60 photos of American politicians and that they would have to guess the political affiliation of each one. They were also informed that only Democrats and Republicans would be shown, and that they would be presented with all male or female candidates (depending on the gender condition they were assigned to).

After they completed the 60 trials, participants were asked a series of demographic questions. In addition to their age and gender, they were asked which political party they most strongly identified with (they could select from four options: "Democrat", "Republican", "Other", and "None"), whether they had American citizenship, whether they had ever lived in the United States, whether they currently lived in the United States, and whether they had voted in any American elections. Following these questions, they were asked to indicate whether they recognized any of the politicians they had been shown in the study. Those participants who indicated that they recognized at least one of the politicians in the study were then asked to try to recall the name, political affiliation, and election state for the candidate they best recognized. Finally, participants were asked whether they had participated in the study before. Once they finished responding to these questions, they were taken to a final webpage that informed them of their overall performance (how many candidates, out of 60, they correctly categorized) and thanked them for their participation.

Results

Our dependent measure of interest was the judgment accuracy achieved by each participant. This proportion was calculated by dividing the number of correct judgments that a participant made by 60 (the total number of trials). Participant political affiliation was not associated with accuracy: $F(3, 1001) < .3$. Nor, for that matter, were any of the binary demographic variables: all ts $< .8$. Participant age did not correlate with accuracy either: $r(1002) = .02$, *ns*. Participants who indicated that they recognized at least one of the politicians in the experiment were not significantly more accurate than those who reported recognizing none of the politicians: 57% vs. 56%, $t(1003) < .5$. Finally,

participants were just as accurate, whether they were assigned to male or female political candidates (accuracy = 56% in both conditions, $t(1003) < .2$). Since none of these factors seem related to judgment accuracy, we collapsed across them and focused on our two main variables of interest: the proportion of Democrats shown in the experiment (i.e., the base-rate) and whether participants were informed of this base-rate. Figure 3 shows how mean accuracy-levels vary with these two factors.

Figure 3 about here

First, we see that participants performed significantly better than chance, even for uniform base-rates (i.e., 50% Democrats) or when they did not know the base-rates[3]. This implies that participants were able to draw some useful information from the politicians' photos, and thus to perform above chance when no other cues (such as the base-rates) were available. In our study, the mean accuracy-level achieved by participants assigned to equiprobable base-rates was 55%, regardless of base-rate knowledge or whether we excluded participants who reported recognizing at least one of the politicians[4]. This level of accuracy is comparable to the one reported by Benjamin and Shapiro (in press): They found that naïve participants who were shown 10-second silent video clips of televised gubernatorial election debates between Democratic and Republican candidates were able to identify the Democratic contender in 53% of the clips. Benjamin and Shapiro concluded that their participants were no better than chance at guessing political

affiliation, but their sample size was much smaller than ours and, as a result, they may have lacked the statistical power to detect a better-than-chance performance.

Now we turn to the effects of the base-rates: A 9 (base-rate level) by 2 (base-rate disclosed or not) ANOVA revealed a main effect of base-rates ($F(8, 987) = 6.98$, $p < 6\times10^{-9}$, $\eta^2 = .05$), a main effect of base-rate knowledge ($F(1, 987) = 100.81$, $p < 2\times10^{-22}$, $\eta^2 = .09$), and an interaction between these two variables ($F(8, 987) = 13.07$, $p < 5\times10^{-18}$, $\eta^2 = .10$). As Figure 3 illustrates, accuracy varied with the base-rates, but mainly when these were known. In fact, the effect of base-rates was significant when base-rates were revealed ($F(8, 487) = 14.57$, $p < 4\times10^{-19}$, $\eta^2 = .19$) but only marginally so when they were not ($F(8, 500) = 1.87$, $p = .063$, $\eta^2 = .03$).

To explore the shape of the relationship between base-rates and accuracy, we ran separate regressions for the two base-rate knowledge conditions, while including both a linear term and a quadratic term for the base-rates (the latter term was obtained by squaring the base-rates after subtracting 50% from each one). When the proportion of politicians who were Democrats was known, there was both a negative linear effect ($\beta = -.12$, $t(493) = 2.91$, $p < .004$) and a positive quadratic effect ($\beta = .42$, $t(493) = 10.28$, $p < 2\times10^{-22}$) of base-rates on accuracy. The positive quadratic effect shows that participants who are informed of the base-rates do make use of this information, at least to some extent. In contrast, when the proportion of Democrats was not known, the linear effect was not significant ($\beta = -.03$, $t(506) < .7$), and there was a *negative* quadratic effect of base-rates on accuracy ($\beta = -.12$, $t(506) = 2.62$, $p < .01$). The negative quadratic effect of base-rates on accuracy could simply be a result of Bayes' Theorem: As the base-rates (or prior probabilities) become more extreme, the added diagnostic value of facial cues

diminishes. As a result, judges who are ignorant of the priors and rely solely on appearances to infer political party would see their accuracy diminish (toward chance) as the base-rates diverge from equiprobability.

Although participants who were informed of the base-rates beforehand did seem to take this knowledge into account, Figure 3 shows that they did not make full use of this information. In particular, we can see that our participants performed significantly worse than the dominant base-rates for every level of the base-rates except 50% (i.e., equiprobability).

To examine judges' performance in more detail, we analyzed their responses using a signal detection framework. Specifically, we calculated nonparametric measures[5] of each person's *sensitivity* (the ability to discriminate Democrats from Republicans; also called *A'*) and *response bias* (the tendency to favor guessing that candidates fall into one particular party; also called *B"*), then submitted the resulting values to the same analyzes we carried out on mean accuracy-levels. Figures 4a and 4b show how these two measures vary across conditions.

Figure 4 about here

Turning first to our measure of sensitivity (*A'*), we can see, from Figure 4a, that participants were reliably able to discriminate Democrats from Republicans in nearly every condition (i.e., $A' > .5$). Furthermore, it seems that knowledge of the base-rates did

not improve sensitivity. The 9×2 ANOVA revealed a marginally significant main effect of base-rates ($F(8, 987) = 1.80$, $p = .074$, $\eta^2 = .01$), a main effect of base-rate knowledge ($F(1, 987) = 4.04$, $p < .05$, $\eta^2 = .004$), and an interaction between these two variables ($F(8, 987) = 2.37$, $p < .02$, $\eta^2 = .02$). In particular, the effect of base-rates was significant when these were known to judges ($F(8, 487) = 2.71$, $p < .007$, $\eta^2 = .04$) but not otherwise ($F(8, 500) = 1.10$, *ns*). A regression revealed both a negative linear effect ($\beta = -.14$, $t(493) = 3.04$, $p < .003$) and a negative quadratic effect ($\beta = -.11$, $t(493) = 2.48$, $p < .02$) of base-rates on *sensitivity* for participants who knew the base-rates (neither coefficient was significant when base-rates were unknown). In other words, for participants who knew the base-rates, the ability to discriminate between Democrats and Republicans seemed to worsen as the proportion of Democrats increased, and did so at an increasing rate.

Turning now to our measure of response bias (*B"*), Figure 4b shows that the tendency to classify politicians into one particular party covaried with the base-rates. Interestingly, however, this relationship seems to reverse depending on whether participants were informed of the base-rates or not. The 9×2 ANOVA revealed a main effect of base-rates ($F(8, 987) = 4.19$, $p < 7\times10^{-5}$, $\eta^2 = .03$) and an interaction between base-rate level and base-rate knowledge ($F(8, 987) = 17.45$, $p < 2\times10^{-24}$, $\eta^2 = .12$). The effect of base-rates was significant whether judges knew the base-rates ($F(8, 487) = 13.62$, $p < 8\times10^{-18}$, $\eta^2 = .18$) or not ($F(8, 500) = 4.16$, $p < 8\times10^{-5}$, $\eta^2 = .06$). A pair of regressions revealed a negative linear effect of base-rates for participants who were informed of the base-rates ($\beta = -.39$, $t(493) = 9.45$, $p < 2\times10^{-19}$) and a positive linear effect of base-rates for those who were not ($\beta = .21$, $t(506) = 4.85$, $p < 2\times10^{-6}$) (in neither group were there significant quadratic effects). In other words, judges who were informed

of the base-rates showed a stronger propensity to guess "Democrat", the higher the proportion of Democrats. This result simply shows that participants adjusted their threshold for guessing "Democrat" up or down, depending on whether they knew the proportion of Democrats to be low or high. It also helps explain the convex shape of their accuracy-levels (Figure 3). We also found, unexpectedly, that judges who did not know the base-rates showed the opposite pattern: As the proportion of Democrats increased, they were more biased in favor of guessing "Republican". It is unclear why judges who were ignorant of the base-rates would show a response bias, of any sort, for the extreme base-rates. This bias trend may have further contributed to the concavity of their accuracy-levels (Figure 3).

As with Study 1, there are two reasons to be skeptical of the possibility that participants who knew the base-rates failed to reach dominant base-rate accuracy because they were probability matching. First, the current experimental design did not involve giving participants feedback after each trial concerning the accuracy of their judgments, in contrast to those designs that yield probability matching behavior. Second, the data do not support a probability matching account. Accuracy-levels for participants informed of the base-rates differed significantly from probability matching accuracy for all but two base-rate levels (30% and 70% Democrats), and fell significantly below probability matching accuracy for base-rates below 30% and above 70%. To further determine whether our participants were probability matching, we examined the proportion of times each participant guessed that a politician was a Democrat. Figure 5 illustrates how the mean tendency to guess "Democrat" varied as a function of the base-rates and base-rate knowledge, and how it compares with what we would expect to see from participants

who were probability matching. Although this tendency varied positively with the base-rates when these were known (but not otherwise), it differed significantly from the probability matching tendency at every base-rate level except 50%. Thus, our results are unlikely to be mainly due to participants probability matching.

Figure 5 about here

Discussion

As with Study 1, we found that mean accuracy-levels were nearly always above chance, regardless of the base-rates. In fact, even participants who did not know the base-rates performed above chance (this was also true if we excluded participants who reported recognizing at least one of the candidates). This suggests that people are, to some extent, able to infer a candidate's political affiliation simply from his or her appearance (Jahoda, 1954). But how do appearance-based inferences interact with base-rate information to influence judgments? Accuracy varied with the base-rates when these were known and it increased the more they deviated from uniformity (i.e., from an equal proportion of Democrats and Republicans). Under uniform base-rates (50% Democrats), informing participants of the base-rates had no effect on their accuracy. This suggests that participants who were given the base-rates made use of this information to some extent. However, as Figure 3 reveals, they underutilized the base-rates, and their accuracy-levels were significantly below what they would have achieved by ignoring the photos and

relying only on the base-rates. For participants who knew the base-rates, we can see that, although mean accuracy increased with the size of the dominant base-rate, it did not do so to the same extent as it would have under the optimal guessing strategy or even the suboptimal probability matching strategy. Consider, for example, the fact that participants in the 70% Democrats condition were just as accurate (55% correct, on average), whether they were informed of the base-rates or not, even though knowledge of the base-rates should have led to no less than 70% accuracy. Or consider participants who were informed that there were 80% Democrats in the experiment. On average, they managed to correctly categorize 3 out of 5 targets (or 60%), which is clearly above chance. Yet is this really a laudable performance given that they knew that 4 out of 5 targets would be Democrats? When base-rate information is available and useful (i.e., nonequiprobable), appearances seem to be detrimental cues, overall.

General Discussion

The two large-sample studies reported in this paper show that judges were generally less accurate at predicting other people's characteristics than they would have been, had they simply ignored appearances and relied on available information concerning the underlying distribution of characteristics. In Study 1, we analyzed data from a popular website that allows people to predict specific facts about each other from their photos and found that website users could often improve their guessing accuracy by relying only on their prior beliefs and the feedback they received on each trial. Using a web-based experiment in Study 2, we replicated this result and showed that it occurs because judges underutilize information about the base-rates. These studies demonstrate a

striking failure to properly integrate information from prior beliefs and from photos: A judge with access to the base-rates[6] *and* appearances should perform *at least as well* as one who only has access to one of these cues. Yet we find that appearance cues are generally detrimental to judgment accuracy.

In fact, when base-rates departed substantially from equiprobability, we found not only that accuracy-levels were below the dominant base-rate but that they fell below probability matching levels as well. In Study 1, we found that for 5 out of 11 characteristics, a probability matching strategy would have significantly outperformed website judgments. The 5 characteristics in question included those with the 4 highest dominant base-rates. In Study 2, we found that accuracy-levels for participants informed of the base-rates fell significantly below probability matching accuracy when the dominant base-rate exceeded 70%. Since probability matching accuracy-levels represent those we would expect from a judge who not only had no access to the photos but also used the base-rates in an inconsistent, suboptimal fashion, this represents a rather low benchmark of accuracy. Nonetheless, judges with access to both the photos and the base-rates often failed to achieve it, further illustrating the negative effects of using appearance cues.

Only in cases where prior beliefs (and feedback) failed to provide useful information --when the dominant base-rates approached chance levels-- did judges seem to benefit from the photos. This suggests that judges overweighed appearances and underutilized base-rates. Our results also show a remarkable level of overconfidence in the ability to infer characteristics from appearances. If one knows the dominant base-rate is X%, then to make a judgment, based on appearances, that goes against the dominant

base-rate is to believe (at least implicitly) that one has enough evidence from appearance cues alone that, if the base-rates were equal, then the probability of accurately judging the characteristic would be greater than X%. Yet, as our studies show, this ability is actually quite low (e.g., producing only 55% accuracy in Study 2). The finding that participants are insensitive to base-rates, especially when they are provided with individuating information about targets, and that they are often overconfident in their ability to draw inferences from the latter type of information is consistent with previous experimental evidence (Dunning, Griffin, Milojkovic, & Ross, 1990; Tversky & Kahneman, 1974). Here we've demonstrated similar effects, for appearances, in a more diverse and much larger sample of respondents, who provided judgments under more naturalistic conditions (i.e., outside the laboratory setting).

These results are noteworthy for their theoretical and practical implications, which run contrary to previous optimistic conclusions regarding the benefit of relying on appearances when drawing inferences about others (e.g., Rule & Ambady, 2008a; Rule, Ambady, & Hallett, in press). This optimism has largely been nourished by studies showing that experimental participants perform above chance when guessing target characteristics from photos, and indeed, we replicate this basic finding in both studies and across all characteristics. But chance levels are rather feeble performance milestones, especially when one or more of the defining categories are highly predominant. In fact, it's safe to assume that the vast majority of human characteristics are non-uniformly distributed across the population in this way. Despite this fact, very few studies have used target category membership frequencies that reflected real world base-rates, with many experimenters opting for equiprobability instead (i.e., targets were equally likely to fall

into each category). As a result, judges could not rely on their prior beliefs and the resulting data had little to say about the relative merits of appearances versus prior beliefs as social judgment cues.

The current studies reveal what happens when researchers do incorporate natural base-rates and prior beliefs into analyses of judgment accuracy. In particular, three notable findings seem to emerge: (1) people are generally better than chance at inferring characteristics from appearances, even when category frequencies naturally approach equiprobability (as they did with the "College degree" variable in Study 1, for example), (2) people's prior beliefs reflect population base-rates quite closely (see Study 1 and footnote 2), and yet, (3) when both appearances *and* prior beliefs are available, judgment accuracy drops *below* what they could achieve by relying on prior beliefs alone. The implication is that appearances, while providing some useful information about target characteristics, have a far lower diagnostic value than our beliefs about the base-rates (i.e., our subjective priors), especially when these base-rates are highly non-uniform. However, when appearance cues are available, people tend to neglect their prior beliefs. As a result, the small benefits provided by appearances (as sources of information) are heavily outweighed by the costs of relying too much on these visual cues and too little on our subjective priors. Only when the base-rates reach chance levels do appearances seem to improve accuracy, which amounts to saying that these latter cues are better than no information at all (i.e., better than chance), but not by much, especially when compared to other, more valid sources of social inference.

In sum, when we consider not just how people utilize appearances in a vacuum, but how they integrate this information with other available cues, a much less flattering

picture emerges. In most real-world contexts, where various social cues are available, reliance on appearances may actually make us worse at predicting the characteristics of others.

Future Directions

We hope future research on appearance-based inferences will move beyond simply demonstrating that people can evaluate others' characteristics from their appearances with above-chance accuracy. As we have argued, showing that first impressions exceed chance levels in a standard lab setting says little about the diagnostic value of appearances in the real world. In fact, the dissemination of such results may have the negative effect of promoting stereotyping based on low-validity visual cues.

More generally, we feel that the field of person perception has run its course in showing that the accuracy of first impressions does (or doesn't) exceed chance. Similarly, the finding that people tend to neglect relevant base-rates when individuating information is available is now well established in the social judgment literature. A more fruitful agenda for the field would be to identify the specific visual cues that people use when they draw inferences from appearances, to measure the diagnostic validity of these various cues, and to distinguish cues that promote accurate social judgments from those that lead judges astray. It would also be interesting to see how these cues differ depending on the context or the specific characteristic being inferred. The next generation of researchers might therefore measure or manipulate various features of appearances, to see how these relate to judgment accuracy. They might also vary the mind-sets or inference goals of judges, in order to examine how these manipulations affect the use of

cues and moderate judgment accuracy.

References

Ambady, N., Hallahan, M., & Conner, B. (1999). Accuracy of judgments of sexual orientation from thin slices of behavior. *Journal of Personality and Social Psychology, 77*, 538-547.

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.

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

Benjamin, D. J., & Shapiro, J. M. (2009). Thin-slice forecasts of gubernatorial elections. *Review of Economics and Statistics, 91*, 523-536.

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

Bull, R., & Hawkes, C. (1982). Judging politicians by their faces. *Political Studies, 30*, 95-101.

Bull, R., Jenkins, M., & Stevens, J. (1983). Evaluations of politicians' faces. *Political Psychology, 4*, 713-716.

Dunning, D., Griffin, D.W., Milojkovic, J.H., & Ross, L. (1990). The overconfidence effect in social prediction. *Journal of Personality and Social Psychology, 58*, 568-581.

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

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review, 112*, 912-931.

Estes, W. K. (1972). Research and theory on the learning of probabilities. *Journal of the American Statistical Association, 67*, 81-102.

Gorn, G. J., Jiang, Y., & Johar, G. V. (2008). Babyfaces, trait inferences, and company evaluations in a public relations crisis. *Journal of Consumer Research, 35*, 36-49.

Hall, C. C, Goren, A., Chaiken, S., & Todorov, A. (2009). Shallow cues with deep effects: Trait judgments from faces and voting decisions. In E. Borgida, J. L. Sullivan, & C. M. Federico (Eds.), *The Political Psychology of Democratic Citizenship* (pp. 73-99). New York: Oxford University Press.

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

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Thousand Oaks, CA: Sage.

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

La Fontaine, J. D. (1668/1974). *Fables (Livres I à VII)*. Paris: Guallimard.

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

Naylor, R. W. (2007). Nonverbal cues-based first impressions: Impression formation through exposure to static images. *Marketing Letters*, *18*, 165–179.

Olivola, C. Y., Eastwick, P. W., Finkel, E. J., Hortaçsu, A., Ariely, D., & Todorov, A. (2009). A picture is worth a thousand inferences: First impressions and mate selection in Internet matchmaking and speed-dating. Working paper. University College London.

Olivola, C. Y., & Todorov, A. (2009, May 5). The look of a winner. *Scientific American*. Retrieved August 1, 2009, from http://www.scientificamerican.com

Olivola, C. Y., & Todorov, A. (In press). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior.*

Pope, D. G., & Sydnor, J. (2008). What's in a picture? Evidence of discrimination from Prosper.com. Working paper. University of Pennsylvania.

Ravina, E. (2008). Love & loans: The effect of beauty and personal characteristics in credit markets. Working paper. Columbia University.

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 of personality from Chief Executive Officers' appearance predict company profits. *Psychological Science, 19,* 109-111.

Rule, N. O., Ambady, N., & Hallett, K. C. (In press). Female sexual orientation is perceived accurately, rapidly, and automatically from the face and its features. *Journal of Experimental Social Psychology*.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*, 137-149.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*, 1623-1626.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (In press). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*.

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124-1131.

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

Zarkadi, T., Wade, K. A., & Stewart, N. (In press). Creating fair lineups for suspects with distinctive features. *Psychological Science*.

Zebrowitz, L. A. (1996). Physical appearance as a basis of stereotyping. In C.N. Macrae, C. Stangor, & M. Hewstone (Eds.). *Stereotypes and stereotyping* (pp. 79-120). New York: Guilford Press.

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., Hall, J. A., Murphy, N. A., & Rhodes, G. (2002). Looking smart and looking good: Facial cues to intelligence and their origins. *Personality and Social Psychology Bulletin*, *28*, 238-249.

Zebrowitz, L.A., & McDonald, S.M. (1991). The impact of litigants' babyfacedness and attractiveness on adjudication in small claims courts. *Law and Human Behavior*, *15*, 603-623.

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

Zebrowitz, L. A., & Rhodes, G. (2004). Sensitivity to "bad genes" and the anomalous face overgeneralization effect: Cue validity, cue utilization, and accuracy in judging intelligence and health. *Journal of Nonverbal Behavior*, *28*, 167-185.

Footnotes

1) By "perceptually ambiguous characteristics", we mean characteristics that cannot be easily inferred from a person's photo, as opposed to ethnicity or gender, for example, which are often easy to judge from appearance. And by "clear categories", we mean those forming discrete subgroups that people can classify themselves into with high confidence, as opposed to variables without clear boundaries or for which people may not be able to reliably classify themselves (e.g., number of hours spent working per week).

2) As an additional measure of how well people's prior beliefs (about the distributions of characteristics) correspond to actual category frequencies, we asked 37 mall shoppers (32% male; age range: 19-60, M = 39.62, SD = 12.92) to estimate the percentage of Americans who fall into the affirmative category for each of the 10 binary characteristics. We found that respondents' estimates correlated well with the base-rates on the website. The mean individual-level correlation (obtained by first correlating each respondent's estimates with the website base-rates, then averaging these correlations across respondents) was 0.58 (p = .0002). The ecological correlation (obtained by first averaging respondents' estimates, then correlating these means with the website base-rates) was 0.85 (p = .002).

3) For participants who were not informed of the base-rates, mean accuracy was significantly above chance for every level of the base-rates except 90% Democrats. For this latter base-rate, the mean accuracy was marginally significantly better than chance

($t(58) = 1.85$, $p = .070$ with a two-tailed test). These results hold even if we only consider the subset of participants who did not report recognizing any of the politicians: Their mean accuracy was significantly above chance for every base-rate level except 10% and 90% Democrats (for 10% Democrats, the mean accuracy was marginally significantly better than chance, $t(53) = 1.70$, $p = .095$ with a two-tailed test).

4) The objective of the earlier studies conducted in the U.K. (Bull et al., 1983; Bull & Hawkes, 1982; Jahoda, 1954) was to test whether perceptions of political affiliation affect personality attributions. The rate of accurate identification of political affiliation (Conservative vs. Labour) was 59.5% in Jahoda (1954), where 50% is chance. This rate cannot be computed in a straightforward fashion from the other two studies because perceptions of political affiliation were measured on continuous scales and the authors classified these ratings into 3 categories: unclassified, perceived as Conservative, and perceived as Labour. According to this classification, in the first study (Bull & Hawkes, 1982), 3 out of 14 politicians were not classified and 7 out of 11 were perceived accurately. This would correspond to a possible range of accuracy from 50% (assuming none of the "unclassified" politicians would have been perceived accurately) to 63.6% (assuming all of the "unclassified" politicians would have been perceived accurately). In the second study (Bull et al., 1983), 13 out of 36 were not classified and 14 out of 23 were perceived accurately, which would correspond to a possible accuracy range from 38.9% to 60.9%.

5) The nonparametric measures of sensitivity and response bias (*A'* and *B"*) require fewer assumptions than the conventional (and parametric) measures (*d'* and β), and are therefore more robust to violations of these assumptions (see Stanislaw & Todorov, 1999). The procedure for obtaining these measures is as follows: Consider only trials in which Democrats were presented and let H_i be the proportion of those trials in which judge *i* (correctly) guessed "Democrat". Now consider only trials in which Republicans were presented and let F_i be the proportion of those trials in which judge *i* (incorrectly) guessed "Democrat". We can calculate judge *i*'s sensitivity (A_i') and response bias (B_i'') using the following equations:

$$A_i' = \begin{cases} \frac{1}{2} + \frac{(H_i - F_i)(1 + H_i - F_i)}{4H_i(1 - F_i)} & \text{when } H_i \geq F_i \\ \frac{1}{2} - \frac{(F_i - H_i)(1 + F_i - H_i)}{4F_i(1 - H_i)} & \text{when } H_i < F_i \end{cases}$$

$$B_i'' = \begin{cases} \frac{H_i(1 - H_i) - F_i(1 - F_i)}{H_i(1 - H_i) + F_i(1 - F_i)} & \text{when } H_i \geq F_i \\ \frac{F_i(1 - F_i) - H_i(1 - H_i)}{F_i(1 - F_i) + H_i(1 - H_i)} & \text{when } H_i < F_i \end{cases}$$

6) To reach dominant base-rate accuracy for a given characteristic, a judge only needs to know which category is the most frequent, not the exact frequency of each category.

Acknowledgments

The authors would like to thank Hart Blanton, Nick Chater, Matthew Salganik, and an anonymous reviewer for helpful comments, Gareth Cook and *Scientific American* for helping us recruit the respondents in Study 2, as well as Julia Hernandez, Valerie Loehr, and Jenny Porter for their excellent research assistance. The authors are especially grateful to Jeff Zemla for programming the "Political Guessing Game" and to the creators of *www.whatsmyimage.com*: Robert Moore, Sameer Shariff, Kevin Shi, and William Macreery, for generously sharing their website's data.

Figure Caption

Fig. 1. An example screen shot from the "What's My Image?" website in Study 1. Photos of real users are replaced with pictures of the authors to maintain anonymity. The larger photo in the middle of the screen represents the target for the current trial. A question about the target appears above his picture, along with the relevant response categories. Feedback about the previous trial is presented, along with the previous target's photo, on the left side of the screen. Advertisements on the right side of the screen are hidden.

Fig. 2. Accuracy-levels under three different strategies for each characteristic (ordered by dominant base-rate): (i) the frequency of the most prevalent categories for each characteristic -or- dominant base-rate accuracy-levels (striped bars), (ii) mean observed accuracy-levels of judges on the website (light grey bars with error bars), and (iii) accuracy-levels achievable by survey respondents (black bars). The words above the bars indicate, for each characteristic, the most frequent response provided by targets (i.e., the dominant category - see also Table 1). Error bars represent 95% confidence intervals.

Fig. 3. Mean accuracy-levels as a function of base-rates and base-rate knowledge (solid black and grey lines with markers). In addition, dominant base-rate accuracy-levels (dashed grey line) and probability matching accuracy-levels (dotted grey line) are shown. Error bars represent 95% confidence intervals.

Fig. 4. Mean sensitivity (*A'* → Fig. 4a) and mean response bias (*B"* → Fig. 4b) as a function of base-rates and base-rate knowledge. The horizontal grey line in each figure represents no sensitivity (Fig. 4a) or no response bias (Fig. 4b). Error bars represent 95% confidence intervals.

Fig. 5. Mean likelihood of guessing "Democrat" as a function of base-rates and base-rate knowledge (solid black and grey lines with markers). In addition, the probability matching likelihoods of guessing "Democrat" are shown (dotted grey line). Error bars represent 95% confidence intervals.

Table 1

Website questions used in statistical analyses and demographic data on selected target users (Study 1).

Characteristic	Question posed on website	# target users	Most frequent category	Proportion of targets who were:		
				Male	In college	Working full-time
Sexual orientation	What is this person's sexual orientation?	654	heterosexual	0.47	0.51	0.46
Use drugs	Does this person use drugs?	608	no	0.48	0.49	0.47
Public high school	Did this person go to public high school?	593	yes	0.47	0.53	0.46
Ever arrested	Has this person ever been arrested?	606	no	0.48	0.50	0.47
Virgin	Is this person a virgin?	596	no	0.48	0.48	0.49
Drink	Does this person drink?	671	yes	0.49	0.50	0.47
Own gun	Does this person own a gun?	148	no	0.40	0.39	0.59
Divorced parents	Are this person's parents divorced?	613	no	0.46	0.50	0.48
Fist fight	Has this person ever gotten into a fist fight?	165	yes	0.38	0.41	0.58
Long term relationship	Is this person in a long term relationship?	649	yes	0.46	0.49	0.48
College degree	Does this person have a college degree?	276	yes	0.54	0.02	0.84

Figure 1

Figure 2

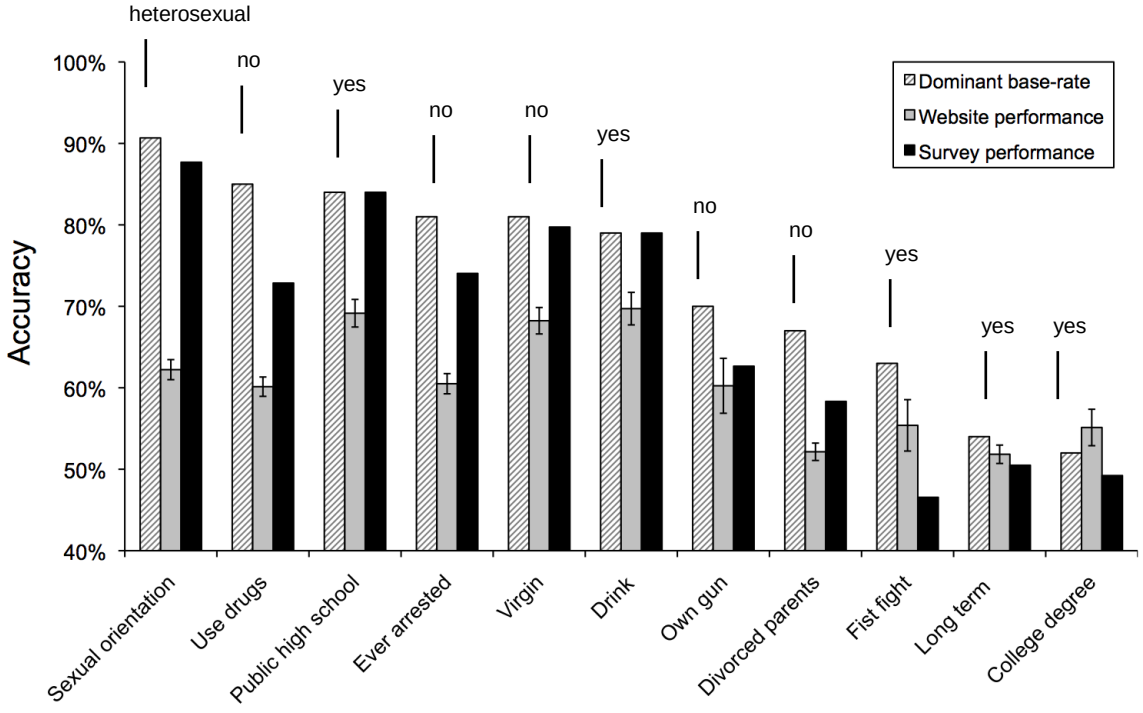

Figure 3

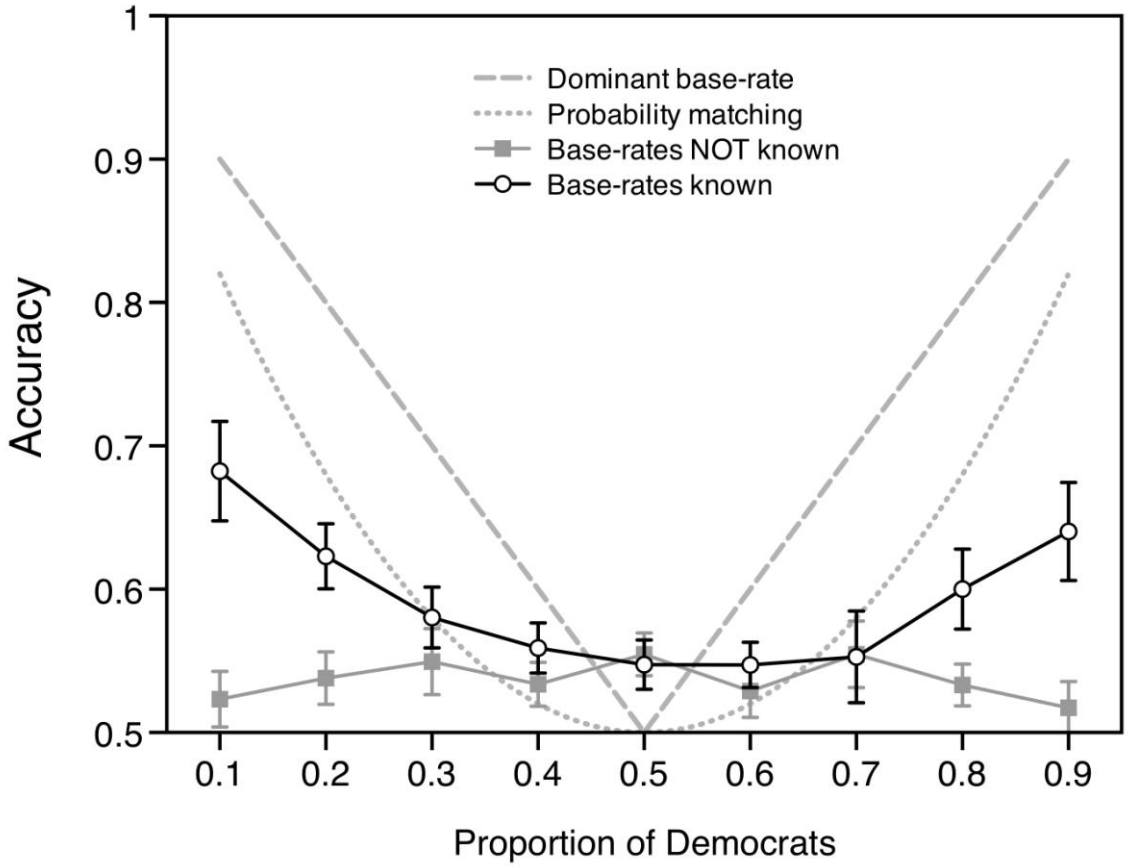

Figure 4a

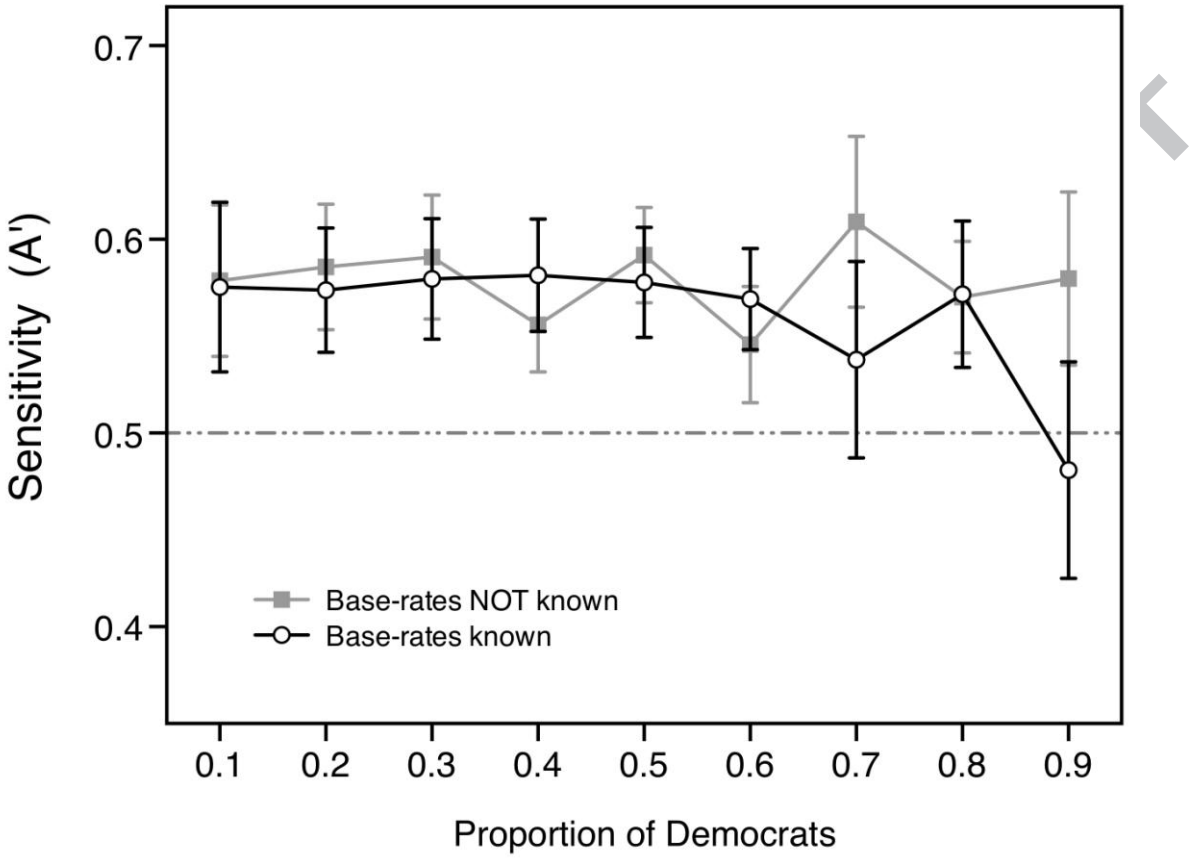

Figure 4b

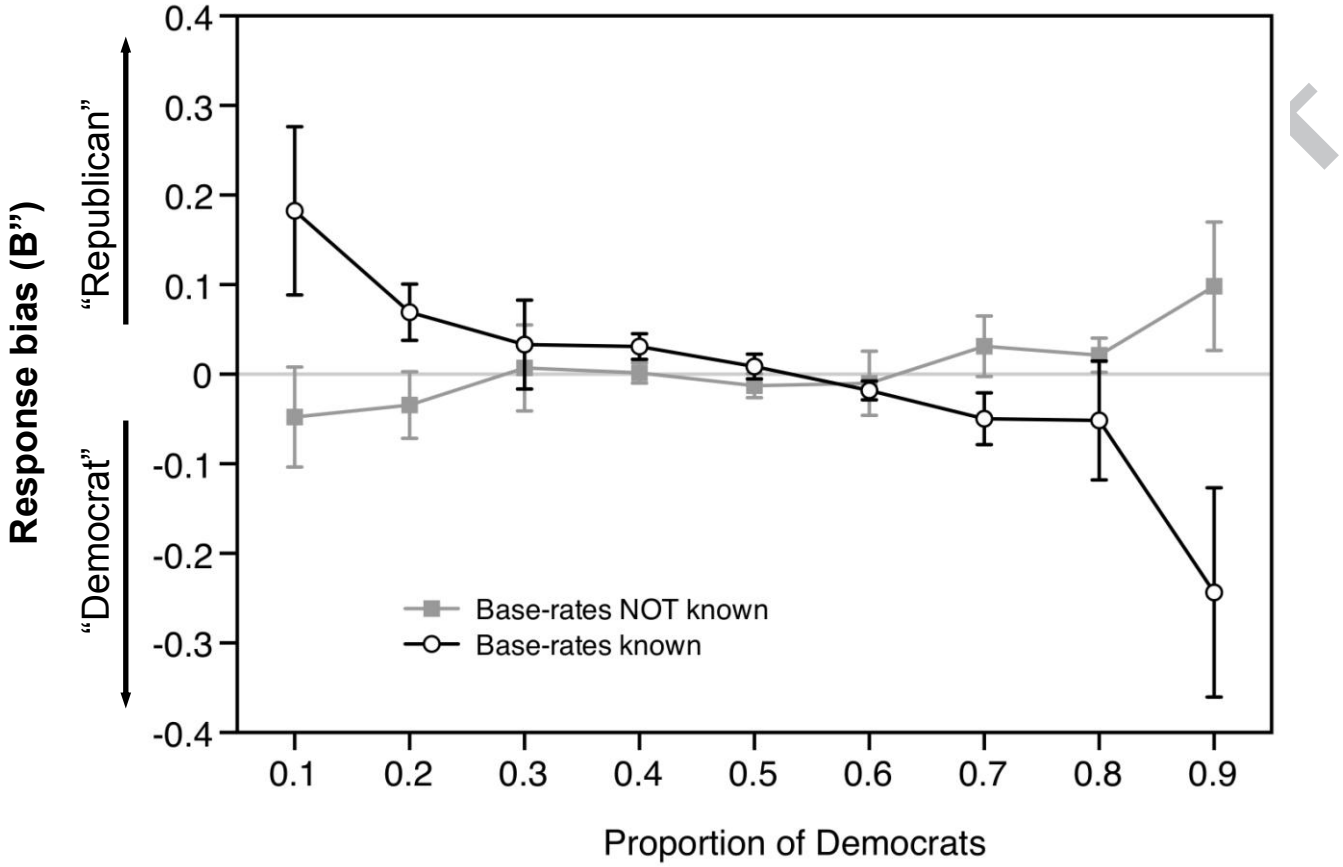

Figure 5