Journal of Experimental Social Psychology 46 (2010) 315–324

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences

Christopher Y. Olivola [a,*], Alexander Todorov [b]

[a] University College London, United Kingdom
[b] Princeton University, United States of America

ARTICLE INFO

Article history:
Received 21 August 2009
Revised 19 November 2009
Available online 15 January 2010

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, 2009; 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, 2009; 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 infor-

* Corresponding author. Address: Cognitive, Perceptual and Brain Sciences, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom. Fax: +44 0 20 7436 4276.
E-mail address: c.olivola@ucl.ac.uk (C.Y. Olivola).

0022-1031/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jesp.2009.12.002

mation 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 (http://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 Fig. 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.

Data set and 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 2000 for a single characteristic. We measured the mean guessing accuracy for each target-characteristic combination. Photos receiving fewer than 10 guesses, for a given characteristic, were discarded from the analysis (but only for that characteristic). 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.

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 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 among 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:

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

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.

Results

We found that male and female targets reported very similar characteristics, the only two major differences being that male

[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).