AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Personality and Social Psychology:
Attitudes and Social Cognition

ISSN: 0022-3514

2025, Vol. 128, No. 1, 61–81
https://doi.org/10.1037/pspa0000422

Learning *Too* Much From *Too* Little: False Face Stereotypes Emerge From a Few Exemplars and Persist via Insufficient Sampling

Xuechunzi Bai[1], Stefan Uddenberg[2], Brandon P. Labbree[3], and Alexander Todorov[4]

[1] Department of Psychology, The University of Chicago
[2] Department of Psychology, University of Illinois Urbana-Champaign
[3] The Institute for Social Research, University of Michigan
[4] Booth School of Business, The University of Chicago

Face stereotypes are prevalent, consequential, yet oftentimes inaccurate. How do false first impressions arise and persist despite counter-evidence? Building on the overgeneralization hypothesis, we propose a domain-general cognitive mechanism: insufficient statistical learning, or Insta-learn. This mechanism posits that humans are quick statistical learners but insufficient samplers. Humans extract statistical regularities from very few exemplars in their immediate context and prematurely decide to stop sampling, creating and perpetuating locally accurate—but globally inaccurate—impressions. Six experiments (N = 1,565) tested this hypothesis using novel pairs of computer-generated faces and social behaviors by fixing the population-level statistics of face–behavior associations to zero (i.e., no relationship). The initial sample contained either 11, five, or three examples with either a positive, zero, or negative linear relationship between facial features and social behaviors. The sampling procedure contained a free-sampling condition in which participants were free to decide when to stop viewing more examples and a fixed-sampling condition in which participants were forced to view all stimuli before making decisions. Consistent with the Insta-learn mechanism, participants learned novel face stereotypes quickly, with as few as three examples, and did not sample enough when they were given the freedom to do so. This domain-general cognitive mechanism provides one plausible origin of false face stereotypes, demonstrating negative consequences when people learn too much from too little.

Statement of Limitations

Although we aim to study a general psychological phenomenon, our ideas and findings are bound by the available literature, methodological choices, and samples of participants. All of these elements add potential subjectivity. First, most references in this article come from the United States and Western Europe. This may constrain the accumulated knowledge to certain historical contexts, which may not be applicable to other contexts. Second, we used tightly controlled experiments because we needed to pin down the precise mechanism and to quantify the proposed causal effect. This design allows for high internal validity at the cost of some external validity. Third, our participants come from one standard crowdsourcing platform; therefore, they possess certain characteristics, such as being a mostly White, English-speaking sample. Although our analysis controlled for individual-level covariates such as age, race, gender, and socioeconomic status, the generalizability of the findings to other samples remains an open question.

Keywords: face stereotypes, statistical learning, sampling, social cognition

Supplemental materials: https://doi.org/10.1037/pspa0000422.supp

Mandy Huetter served as action editor.

All materials, data, experiment code, analytical code, and preregistration materials are anonymized and are publicly available. See supplemental materials for access. Additional supporting information can be found in a separate file. This research was supported by the Richard N. Rosett Faculty Fellowship at the University of Chicago Booth School of Business to A. Todorov. The authors thank members of the Perception and Judgment lab and the Fiske lab for their constructive feedback.

Xuechunzi Bai played a lead role in conceptualization, data curation, formal analysis, methodology, and writing–original draft. Stefan Uddenberg played a supporting role in conceptualization and writing–review and editing and an equal role in data curation and methodology. Brandon P. Labbree played a supporting role in data curation and writing–review and editing and an equal role in software. Alexander Todorov played a lead role in funding acquisition and supervision and an equal role in conceptualization and writing–original draft.

Correspondence concerning this article should be addressed to Xuechunzi Bai, Department of Psychology, The University of Chicago, Kelly Hall 409, 5848 South University Avenue, Chicago, IL 60637, United States. Email: baix@uchicago.edu

> All they have done for me is to make me a little more conscious of how hard it is to classify and to sample, how readily we spread a little butter over the whole universe.
>
> —Lippmann (1922), Public Opinion, p. 104.

Despite the age-old advice not to judge a book by its cover, we struggle to resist this tendency, even though it can lead to inaccurate impressions of people's psychological traits with significant social consequences (see reviews in Todorov, 2017; Zebrowitz, 2017). Why do people engage in appearance–character attributions if they are not necessarily accurate?

The Overgeneralization Hypothesis

One prominent answer lies in the overgeneralization hypothesis (Zebrowitz & Collins, 1997; Zebrowitz, 2017), according to which people form associations between a subset of individuals' appearance and personality traits and then apply the same associations to a broader set of individuals whose appearance resembles the initial subset (Zebrowitz & Collins, 1997). For example, the baby-face overgeneralization states that people pick up associations between babies' facial features (e.g., round faces, large eyes) and personality traits (e.g., submissiveness) and generalize those associations to adults whose facial appearance resembles that of a baby regardless of their actual age (Montepare & Zebrowitz, 1998). Similar accounts have been proposed for other attributions. According to the unfit-face overgeneralization, people overgeneralize associations between facial anomalies and physical or intellectual disability (Zebrowitz et al., 2003). According to the familiar-face overgeneralization, people overgeneralize associations between affective associations and familiar faces (Verosky & Todorov, 2010, 2013; Zebrowitz et al., 2007). And according to the emotion overgeneralization, people overgeneralize associations of emotional states (e.g., happy-looking and approachable/helpful) to emotionally neutral faces resembling the respective emotion (Albohn & Adams, 2020; Engell et al., 2010; Oosterhof & Todorov, 2008, 2009; Said et al., 2009; Zebrowitz et al., 2010).

Although the overgeneralization hypothesis provides a pioneering account of why certain appearance–character attributions are formed and maintained, its underlying mechanism is yet to be established. First, the current theory is domain-specific. It is possible that particular social cues and associations are picked up due to evolutionary pressures, but the process under which people notice and persist in using these cues can be domain-general. We propose a general cognitive mechanism that may underlie the existing overgeneralization effects in such varied domains as babyfacedness, fitness, familiarity, and emotional expressions. Second, even if some associations are rapidly acquired because of evolutionary pressures (Zebrowitz, 2004), not all face stereotypes are evolutionarily justifiable, and some are created with explicitly racist motivations (e.g., the physiognomic theories popular in the 19th and early 20th centuries; see a historical review in Todorov, 2017, Ch. 1). Our proposed mechanism does not require evolutionary justification; it works with any arbitrary associations between appearance and personality traits. Third, the overgeneralization account focuses on the consequences of generalizing prior knowledge to new faces but not the intermediate processes that give rise to this prior knowledge (Bjornsdottir et al., 2024; Cone et al., 2017; Shen & Ferguson, 2021; Todorov & Uleman, 2002; Zebrowitz et al., 2003). Here, we focus on the intermediate processes: How quickly do people learn associations between facial features and personality characteristics? Do people persist in using the learned associations despite counter-examples, and if they do persist, why? Here, we introduce a domain-general cognitive mechanism for the origin of false face stereotypes: *In*sufficient *Sta*tistical *Learn*ing, or Insta-learn for short.

Insufficient Statistical Learning: Insta-Learn

In line with the statistical definition of overgeneralization (generalization error; Hastie et al., 2009) and the overgeneralization hypothesis of face stereotypes (Zebrowitz & Collins, 1997), Insta-learn defines face stereotypes as spurious associations learned from facial appearance and behaviors, implicating personality traits. These associations represent temporary patterns derived from small samples. They are spurious to the extent that they do not accurately reflect patterns in the larger population. Insta-learn proposes that people learn accurately from small samples but stop sampling too early to collect adequate evidence and, hence, allow for spurious associations to persist. As a result, people stick to such locally accurate—but globally inaccurate—face stereotypes. Insta-learn builds on two assumptions from the statistical learning literature: Humans are quick statistical learners but inadequate samplers.

First, people are skilled at extracting statistical regularities from the available social environment in domains such as language acquisition (Saffran et al., 1996), concept learning (Tenenbaum et al., 2011), and social categories (Liberman et al., 2017; Rhodes & Baron, 2019). In the social domain, adult and preschooler participants can infer the preferences of new individuals by observing only minimal statistical information (Vélez & Gweon, 2020). When asked to evaluate human faces, adult participants can extract the underlying statistical distributions of novel faces. In this paradigm, novel faces are generated from various statistical distributions (e.g., Gaussian or uniform). Participants, who implicitly learned the face distribution, tended to evaluate faces closer to the central tendency more positively than faces toward the edges of the distribution (Dotsch et al., 2016). In another study, participants learned novel face stereotypes by observing the relationship between the sellion width of a face and behaviors from as few as 20 training examples (Chua & Freeman, 2022).

Second, the immediate context from which people learn is not always static or exogenous. People not only passively observe the context but also actively, and oftentimes inadequately, create the context (Bai et al., 2022; Denrell, 2005; Fiedler, 2000). Although the evidence for when people oversample or undersample is mixed (Evans et al., 2019), in the domain of forming impressions, research finds that people tend to insufficiently sample information. For instance, people sample much less information than they originally thought they would (Klein & O'Brien, 2018). When evaluating whether they liked a painting style, for example, people originally thought they would choose to view 16 paintings on average before making up their minds, but they only viewed three paintings. Similarly, when forming impressions, participants requested substantially smaller samples than experimenter-determined samples (Prager et al., 2018). Empirically, participants sampled about seven to eight traits from a pool of 36 before making judgments, especially when the initial traits contained negative or extreme signals. Humans are insufficient samplers, which may prevent them from learning the true appearance-character associations.

The combination of being fast statistical learners and insufficient samplers generates locally accurate but globally inaccurate impressions, constituting a plausible candidate for the mechanism of forming false face stereotypes. We highlight four contributions of this account. First, this mechanism is general across domains; it is consistent with various overgeneralization accounts (e.g., fitness, familiarity), but it expands the domain to arbitrary associations as long as the initial samples contain detectable statistical regularities. Second, this mechanism predicts very little training data are sufficient to afford statistical learning; much smaller than prior work has assumed (120 training faces in Zebrowitz et al., 2003; 20 training faces in Chua & Freeman, 2022; 18 training faces in Hill et al., 1990). Third, this mechanism specifies insufficient sampling as the causal effect. In other words, when two people receive the same initial samples indicating identical spurious associations, those who continue to sample will be less likely to form false face stereotypes than those who do not sample sufficiently. Insufficient sampling might have been assumed to contribute to the persistence of face stereotypes by prior researchers (Hill et al., 1990; Todorov et al., 2015), but its causal role has not been empirically validated. Fourth, Insta-learn suggests interventions to change erroneous first impressions at the level of the representational structure, including both crafting diverse initial samples and encouraging sufficient sampling, which complement individual-level strategies.

Contextualize Insta-Learn: Initial Samples and Subsequent Sampling

Multiple mechanisms potentially explain why people do not sample enough. In addition to group-serving motivations or cognitive biases (Fiske, 1998; Fiske & Taylor, 1984; Pratto et al., 1994; Sherman et al., 2000; Turner et al., 1979), two mechanisms of active sampling are most relevant to our work. One mechanism is experience-based feedback (Bai et al., 2022; Denrell, 2005; Le Mens & Denrell, 2011; Rich & Gureckis, 2018). The hot-stove effect posits that people stop interacting with certain groups when they encounter negative experiences with them, suggesting feedback as the key mechanism (Denrell, 2005). The explore-exploit trade-off in stereotype formation posits that people do not sample for more when they are satisfied with the existing groups given prior positive interactions, again assuming interaction feedback as the mechanism (Bai et al., 2022). Another mechanism is sample size (Bott & Meiser, 2020; Prager et al., 2018). Small samples tend to covary with clear-cut and conflict-free impressions, which can prevent people from sampling for more (Prager et al., 2018). Similarly, the pseudocontingency hypothesis posits that the asymmetric size of two groups can create false impressions that the two groups differ even if the underlying ratio is identical (Bott & Meiser, 2020; Hamilton & Gifford, 1976; Meiser & Hewstone, 2010). From this perspective, the difference in sample size, called base rate difference, is necessary for inaccurate impressions to emerge. Although feedback and sample size can be important mechanisms to understand why people sample insufficiently, here we highlight another possibility: The strong statistical regularities in the initial sample can perpetuate inaccurate impressions without needing feedback or varying the size of the sample.

Statistical regularities in the initial sample underlying face stereotypes may emerge under two kinds of cultural contexts: On the one extreme, people who live in a shared cultural context will have similar exposures to strong correlations between facial features and behaviors. The strong correlations may come from media broadcasts or historical affordances such as a skewed sample with more African Americans' appearance relating to crime activities (Grunwald et al., 2022; Lum & Isaac, 2016). Various cultural instruments expose people to strong correlations between facial appearances and characters, leading to shared but inaccurate impressions (Over & Cook, 2018). On the other extreme, people may also have idiosyncratic exposures to varying levels of correlations between facial features and behaviors. Heterogeneous encounters create individual variations when it comes to complex judgments from facial features (Albohn et al., 2022). There is growing evidence suggesting that not all face stereotypes are shared, but some are more idiosyncratic (Martinez & Todorov, 2023). We hypothesize the real world sits between the two extremes, and the following set of studies provides an experimental approach to empirically examine the consequences of the initial statistics in the sample when they are shared (Experiments 1–4 and 6, below) versus idiosyncratic (Experiment 5, below).

An Experimental Approach

There are two main challenges in testing the Insta-learn hypothesis. The first challenge lies in disentangling whether a face stereotype is learned from sample or population statistics. In real-world data, sample and population statistics might show similar patterns, perhaps due to self-fulfilling prophecies or explicit discrimination. To solve this problem, we created novel pairings of facial features and social behaviors, pairings for which participants could not possibly have any prior knowledge. This allowed us to experimentally manipulate both the population and sample statistics to test whether participants learn spurious associations from initial samples even though the population statistics would indicate an entirely different relationship. The second challenge lies in identifying whether the spurious association is driven by inadequate sampling on the part of participants or by other cognitive processes. To address this challenge, we made the sampling procedure flexible. We experimentally encouraged participants to flexibly sample more or fewer examples so that we could compare their results with those of participants who were forced to view all the examples in the population.

Building on prior work in experimentally testing the sampling of personality (Prager et al., 2018), Insta-learn can be viewed as a stress test on the mechanism. The input space consists of face images that are high-dimensional and continuous. It is therefore more challenging to learn than the standard text-based personality traits, which are lower dimensional and discrete. Moreover, the personality traits in Insta-learn are not explicit (e.g., a list of adjectives), but rather the traits need to be extrapolated by observing the covariation between face images and behaviors. Incorporating high-dimensional and continuous input space and letting participants extrapolate the underlying traits more truthfully reflect how people may create false impressions in the domain of face perceptions in the real world.

Specifically, participants were asked to make judgments on how generous a new face is based on the participants' past observations of other faces. In all experiments, we showed participants a small sample of computer-generated faces varying along one nonsocial appearance dimension and along generosity (namely, their charitable donation decisions). We then asked participants to either (a) continue viewing all remaining sets of faces and behaviors or (b) stop viewing more faces at any time if they felt confident to make a decision.

We finally asked participants to make judgments on new faces they never saw during the training phase by asking how much they thought this new person would donate (see Figure 1). At the end of each experiment, we collected standard demographic information.

Importantly, unbeknownst to participants, facial appearance and charitable behaviors were completely uncorrelated in the population data set from which the samples were drawn. That is, any given face was equally likely to commit any given behavior. If participants sample all faces and learn the (lack of) covariation accurately during the training phase, the optimal strategy that minimizes expected error is to donate an average amount for each testing trial. However, if participants sample only a few faces and jump to conclusions too soon, they will make decisions according to the initial samples they happened to see. Critically, the initial samples were chosen by the experimenters to show a linear relationship between facial appearance and charitable behavior—a fact also unknown to participants. Hence, if participants make decisions based solely on the initial sample statistics, their decisions should reveal a systematic linear relationship and should not look like random guesses or guesses closer to the mean of the observed contributions.

Experimental stimuli were designed with the following considerations. First, the face stimuli drew on previous research in face perception and impression formation (Todorov et al., 2015; Todorov & Oh, 2021). To avoid prior beliefs on what types of faces look trustworthy or generous, using FaceGen3.1, we generated novel faces that varied in a manner orthogonal to seven core social dimensions: attractiveness, competence, dominance, extroversion, likability, threat, and trustworthiness. Visual inspection of the changing dimension reveals that it roughly reflects the width of a face. The faces were manipulated from −10 to +10 *SD* (−8 to +8 in Experiments 2, 3, 4, 5, and 6), in increments of 2 *SD*. To avoid memory biases, we generated 200 distinctive identities so that participants never saw two identical faces during the task. In the paradigm, participants read how much money a person (exemplified by a face) donated in a trust game or a hypothetical charitable event. The amount indicated levels of the generosity of the stimulus face. The amount of money ranged from $0 to $100, in increments of $10 ($25 in Experiments 2, 3, 4, 5, and 6).

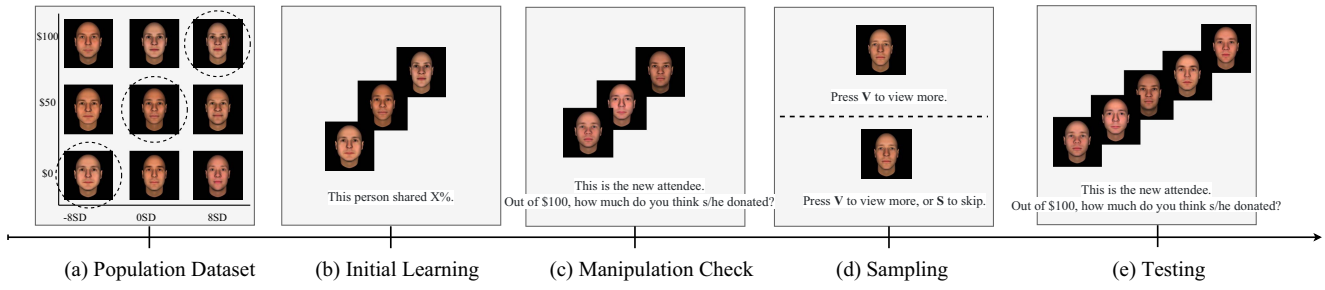
Overview of Studies

Within this paradigm, Experiment 1 manipulated the initial samples of faces and behaviors to be linearly and positively correlated. Half of the participants were given the opportunity to sample as many faces as they would like (i.e., free sampling), and half were asked to sample all faces (i.e., fixed sampling). We hypothesized that participants in the free sampling condition would reveal a more positive linear association in their decisions than participants in the fixed sampling condition. Experiment 2 tested the robustness of the initial sample manipulations by including varying relationships: positive linear, negative linear, and zero association. We hypothesized a main effect of the initial samples and sampling condition differences as in Experiment 1. Experiment 3 tested the hypothesis in a more naturalistic way by removing manipulation checks and fixed sampling conditions. We hypothesized that the results would replicate Experiment 2's free sampling condition. Experiment 4 tested this paradigm with a more stringent design using stronger incentives for accuracy and aggressively fewer trials of initial samples. We hypothesized that the results would replicate Experiment 3. Experiment 5 explored the effects of the differential strength of the initial sample statistics and hypothesized that participants would persist in using spurious associations regardless of weaker initial signals. Experiment 6 explored the utility of additional and reliable information. Participants were given the opportunity to gather additional information. We hypothesized participants should be able to ignore the spurious associations if they utilize additional and reliable information. Finally, taking stock of empirical evidence from the experiments above, we compared participants' behaviors to a set of simulated benchmarks by Bayesian agents to explore the question if participants indeed *learned too much* from *sampling too little*.

We preregistered all hypotheses (except for the mechanism analyses and the benchmark; they were responses to reviews),

Figure 1

A Conceptual Illustration of the Experimental Paradigm



Note. A conceptual schema of the experimental paradigm. (a) An example population training data set with the three initially shown samples highlighted via dotted circles. Sample face stimuli varying along a single compound nonsocial dimension at −8 *SD*, the *M*, and +8 *SD*. The horizontal axis represents face stimuli, and the vertical axis represents the amounts of charitable donations. The population correlation between facial features and donations is zero, but it is not revealed to participants. The sample correlation is positive and linear. (b) The initial learning shows the three face-donation examples. (c) Participants are asked to make initial decisions on new faces as the manipulation check. (d) The remaining face-donation stimuli in the population data set are used for the critical sampling experiment. In this phase, participants are randomly assigned to either fixed or free-sampling experimental conditions. According to the instructions, participants in the fixed sampling condition could "press V to view more" until the end, whereas participants in the free sampling condition could "press V to view more or S to skip to the decision page." (e) Finally, participants are asked to make more decisions on new test faces. See the online article for the color version of this figure.

analysis plans, and study materials at https://osf.io/syc6b/?view_only=555262392d9b4943b2145c0ed00efd23. See details and a live experiment demo via Supplemental Materials. The Institutional Review Board for Human Subjects at Princeton University (protocol number 8,305) approved these studies. All participants gave their consent before the experiment and read debriefing materials after the experiment.

Experiment 1: Establishing Insufficient Statistical Learning

The first study established a baseline paradigm upon which the other experiments built. We predicted that participants would learn and retain linear associations between novel faces and social behaviors even from a population data set that showed no such associations. These associations were formed in two steps. First, an initial sample, biased and small, allowed people to learn a spurious association. We hypothesized that (a) participants would be able to learn the linear relationship from the initial sample, and (b) insufficient sampling would prevent them from unlearning the spurious association. We hypothesized participants in the free sampling condition would be less likely to revise their prior spurious beliefs as they would sample fewer examples than those who were forced to sample the whole population data set. The hypothesis would fail if participants did not demonstrate a linear relationship in their manipulation check decisions or if responses in their final decisions between the two sampling conditions did not differ systematically.

Method

Participants

We recruited 120 participants (7,920 nested trials) from the university subject pool in order to obtain .80 statistical power to detect small- to medium-sized effects in a multitrial between-participants design with two conditions. We intentionally chose subject pool participants because we reasoned that students might be more focused on long tasks than online workers. Although we preregistered to collect data for three initial conditions with a planned size of 300, COVID-19 pandemic prevented us from continuing data collection after the first condition. Hence, the data collection stopped at the current sample size. After modifications to the online platform, we were able to collect all three conditions as planned (see below).

Procedure

Participants received an online experiment link via email. They were instructed to complete the study within 24 hr and to provide a completion code to receive credit. The completion time was around 45 min. The experimental procedures can be divided into four phases: initial sample learning, manipulation check, sampling, and testing phases (Figure 1).

Participants first agreed to the consent form and to participate all the way to the end of the experiment, via a procedure used to reduce subject attrition in online subject pools. Participants who did not agree to these terms were directed out of the study.

Participants were then provided with general task instructions. Their task was to decide how much money to share with new people. They were first taught the basic rules of a trust game (Berg et al., 1995). This game is often used to test cooperative motives. In the context of the game, participants typically share more monetary units (MUs) with partners they perceive as generous. Specifically, we paired our participants with hypothetical partners instantiated by artificial faces. Participants were given 10 MUs. They were told that they must send some amount of their MUs to an anonymous partner. They could send as few as 0 MU to as many as 10 MUs. They also learned that whatever amount they send to the partner would be quadrupled. The partner would then receive the money and make a similar choice, deciding whether to give some amount of the now-quadrupled money back to the participant. Participants were told that the goal was to earn as many MUs as possible because these units would be converted into a real bonus at the end of the experiment (100 MU = $1). To make the rules easier to understand, we explicitly mentioned that the optimal strategy would be to share more MUs with those who are more likely to send MUs back. We told participants that how much each partner would send back was predetermined and that they could learn about the group of potential partners by observing some of their behaviors first.

Next, participants entered the initial sample learning phase. They saw 11 pairs of faces and past behaviors. We told them that to help with their decisions, they would first see 11 randomly selected pictures of partners and their past donations. The 11 examples were chosen by the experimenters to create a positive linear relation between the appearance of faces and behaviors. The varying facial features are rich and holistic, but roughly, narrower faces donated more ($r = 1.0$). Each stimulus pair of face–behavior was displayed sequentially, with one face image in the center of the screen and one behavioral statement (boldfaced text "This person shared x%") under the face image. Below the behavioral statement was the instruction to "Press V to view next." After pressing V on their keyboard, participants were shown a centered black fixation cross for 50 milliseconds before the next stimulus trial. The faces were presented according to the MUs they returned. For example, the face stimulus that returned 0% of their MUs was shown first, then the face stimulus that returned 10%, and so on (in increments of 10%) until the last face stimulus that returned 100%.

After the initial sample learning phase, participants entered the manipulation check phase. They were asked to share money with 33 new partners from the same group using a slider ranging from 0 to 10 MUs. This was done to make sure participants learned the (intentionally obvious) statistical pattern between faces and their behaviors in the initial small sample. Face stimuli appeared at the center of the screen sequentially and in a random order separately generated for each participant. Under the face image, participants could move the slider to indicate their decisions. They could then advance to the next face.

After the manipulation check phase, participants entered the sampling phase. They were told that there were 110 more members in the group. To encourage accuracy, the following text was displayed in bold: "Decisions based on small samples are not always accurate." Participants could view more examples before making their final decisions. Approximately half of the participants were randomly assigned to the free sampling condition ($N = 62$), while the remainder were assigned to the fixed sampling condition ($N = 58$). In the free sampling condition, participants had the option to view more examples by pressing the V-key or to skip to the decision page by pressing the S-key. In the fixed sampling condition, participants could only view more examples (by pressing the V-key) until they viewed

all faces (i.e., the population data set). As mentioned above—and in contrast to the initial sample—the statistical pattern in the population data set showed no linear relationship at all; variation along the (approximately facial width) dimension was uncorrelated with donation amount ($r = 0.0$).

Finally, participants entered the testing phase. They were asked to share MUs with 66 members of the group (i.e., stimulus faces). As in the manipulation check phase, they could indicate their decisions by moving the slider below the face image anywhere between 0 and 10 MUs. Among the 66 faces, 33 were from the manipulation check phase, and 33 were completely new faces that participants had never seen prior to this phase. And just as in the manipulation check phase, face stimuli appeared at the center of the screen sequentially and in a random order separately generated for each participant. Participants advanced to the next face by clicking a "next" button on the same page. In this testing phase, we tested whether and how participants (a) changed their prior decisions and (b) generalized the learned association to new members.

Results

As predicted, data from the manipulation check phase showed participants learned the positive linear relationship from the initial sample. Using a multilevel model with initial decisions as the dependent variable and faces as the independent variable, with error clustered at the individual level, we observed that participants' decisions were associated with the manipulated facial dimension: For each unit increase in the manipulated facial dimension, participants shared 1.88 more monetary units, $b = 1.88$, 95% CI [1.72, 2.04], $p < .001$ (Figure 2a).

Moreover, data from the testing phase showed different associations were learned as a function of the sampling condition. Participants' decisions in the free sampling condition were more likely to reflect the linear relationship between appearance and behaviors in the initial biased sample than participants' decisions in the fixed sampling condition. In a multilevel model with final decisions as the dependent variable, faces and sampling conditions (0 for fixed and 1 for free) as independent variables, and error clustered at the individual level, we observed that sampling condition significantly interacts with appearance, interaction $b = .32$, 95% CI [0.09, 0.55], $p = .006$ (Figure 2e). Last, we found participants in the free sampling condition on average viewed less than one fifth (23/121) of the total population of faces (Figure 3a).

In sum, we found that participants accurately learned the statistical relationship between novel faces and charitable behaviors from small samples. However, participants in the free sampling condition sampled very few faces and behaviors before making decisions. As a consequence, they were less likely to revise their initial judgments, leading to maintaining incorrect impressions about the association between appearance and behavior.

Experiment 2: Testing Different Statistical Relationships

The second study aimed to rule out an alternative interpretation of the results of Experiment 1. Participants could have applied some metapriors, such as unmanipulated signals in facial features that were accidentally consistent with the positive linearity, regardless of what the initial sample intended to show. For example, faces with greater width are perceived as less trustworthy and less likely to reciprocate (Stirrat & Perrett, 2010), which is consistent with our previous results.

To further establish that participants do learn from the initial samples, we manipulated the statistical relationship in the initial sample. In addition to the positive linear relationship between appearance and behaviors in Experiment 1, we added one condition where the initial sample included the completely opposite relationship (i.e., negative linear) and another condition where the initial sample included no relationship at all (i.e., a zero correlation). We hypothesized that manipulation check tests would reveal a difference in decisions across initial sample conditions. The hypothesis would fail if participants in all three conditions uniformly revealed a positive linear association in their manipulation check decisions.

Method

Participants

We collected data from 320 participants (9,600 nested trials) via Amazon Mechanical Turk Prime (i.e., Now Cloud Research—a service making it easier to recruit and manage higher quality participants on Amazon Mechanical Turk) to ensure the inclusion of at least 50 participants in each condition to detect small- to medium-sized effects in a multitrial between-participants design. We changed our participant pool to online workers due to COVID-19-related restrictions. All participants accessed the link and provided a completion code for monetary compensation. Our participants were 61% male, 75% White, with an average age of 37 and a *SD* of 20, and 42% having earned a bachelor's degree. In addition, we asked for free comments at the end of the experiment; the majority of the participants were satisfied and engaged in the task.

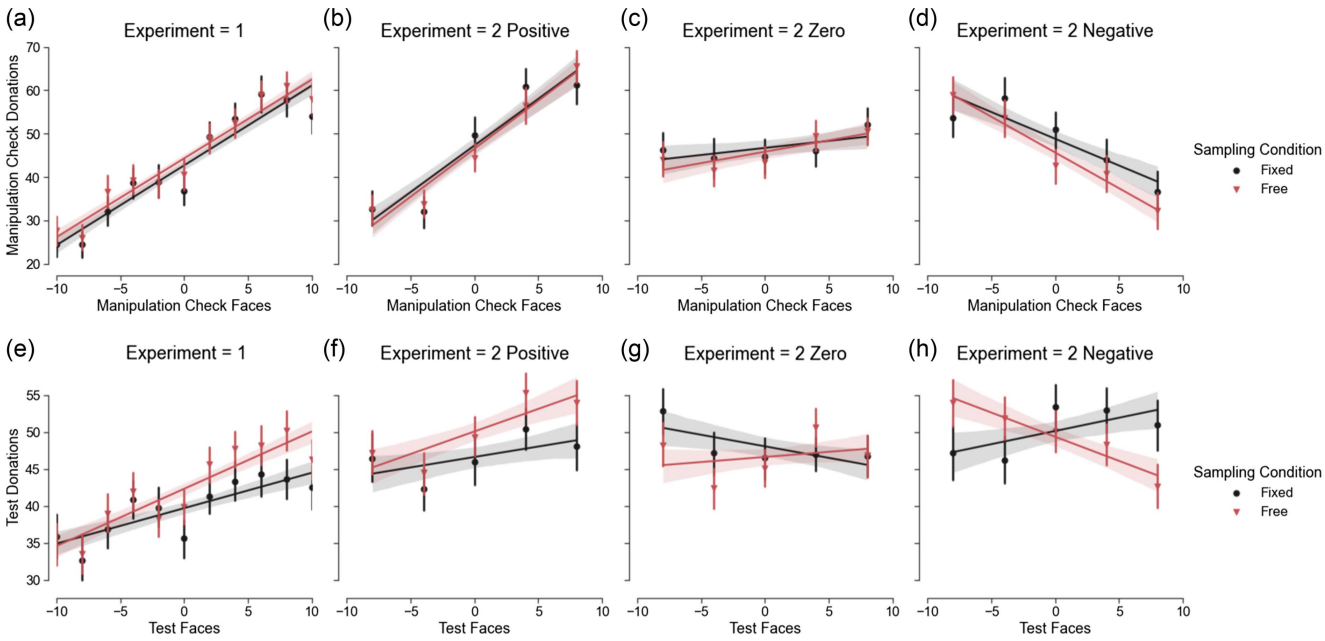
Procedure

The procedures were identical to Experiment 1, with minor changes for simplification purposes. First, we reduced the number of stimuli to ensure response quality from online workers (Robinson et al., 2019). This task took about 10 min with fewer trials and simpler instructions. Second, we simplified the general instructions, replacing the relatively complex trust game with a simpler estimation task. Participants were introduced to a hypothetical charitable event. Their task was to guess how generous a new attendee is based on observations of past attendees. They read, "A large group of people attended a charity event. Each attendee was endowed with $100 and donated some of this amount. Your task is to figure out whether appearance predicts charitable behavior; donating more money." As in Experiment 1, there were four phases.

In the initial sample learning, participants were randomly assigned to see five examples (instead of 11) with either positive ($r = 1.0$, $N = 106$), negative ($r = -1.0$, $N = 100$), or near zero ($r = 0.1$, $N = 114$) correlations. In the manipulation check phase, participants estimated the donations of 15 new examples (instead of 33) from $0 to $100 (instead of 0–10 MUs). In the sampling phase, participants were reminded that they only saw five examples, which may or may not be representative of how the rest of the attendees donated. Participants in the free sampling condition saw boldfaced instructions saying, "You can sample up to 20 additional samples. That is, you can continue viewing more examples (by pressing V-key), or you can stop at any time to make your decisions (by pressing S-key)." In contrast,

Figure 2

Donation Decisions Before (i.e., Manipulation Check) and After (i.e., Testing Phase) the Free Versus Fixed Sampling Conditions in Experiments 1 and 2



Note. Interaction effects of the sampling manipulation from Experiments 1 and 2. The horizontal axis indicates the intensity of the (arbitrary nonsocial) dimension along which test faces varied, ranging from −8 *SD* to +8 *SD* (−10 *SD* to +10 *SD*) in Experiment 1. The vertical axis indicates estimated amounts of charitable giving for a given face, ranging from $0 to $100 (0 MUs to 10 MUs in Experiment 1). Figures are grouped by Experiment 1 (a, e), Experiment 2 positive initial (b, f), Experiment 2 zero initial (c, g), and Experiment 2 negative initial correlation (d, h), with fixed (black dot) and free (red triangle) sampling conditions. The plots represent fitted linear regressions displaying the central tendency and 95% confidence intervals. The upper figures (a–d) represent interaction effects before the sampling manipulation. As expected, participants in both conditions saw the same initial examples, so there should be no differences. The lower figures (e–h) represent interaction effects after the sampling manipulation. As hypothesized, participants in the free, more than in the fixed, sampling condition were more likely to use the associations learned from the initial samples, that is, steeper slopes in red triangle lines. MU = monetary unit. See the online article for the color version of this figure.

participants in the fixed sampling condition saw boldfaced instructions saying, "You will see 20 additional examples. That is, you will continue viewing more examples before making your decisions (by pressing V-key only)." In the testing phase, participants evaluated 30 new faces (instead of 66), with 15 old faces from the manipulation check and 15 new faces they had never seen before. At the end of the study, standard demographic information was collected.

Results

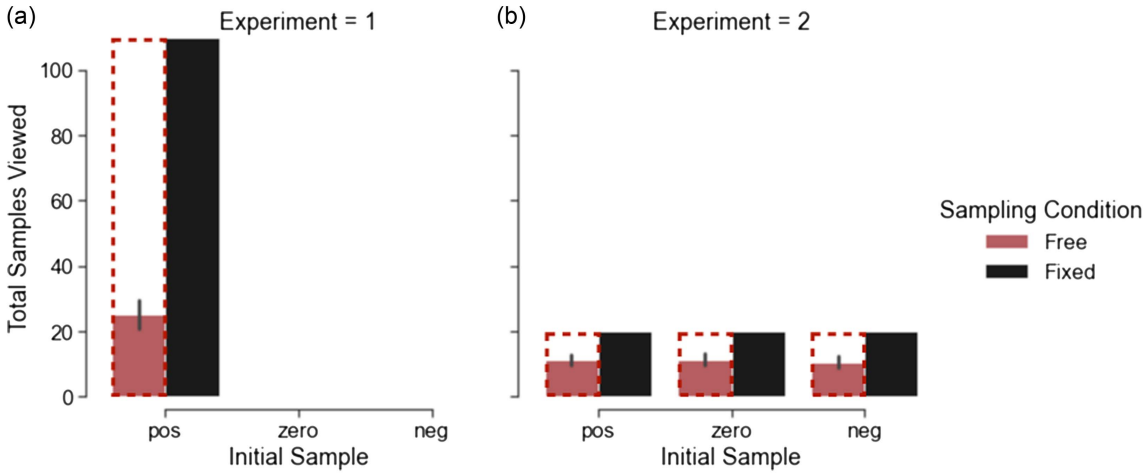
As predicted (Figures 2b, 2c, 2d), data from the manipulation check phase showed that participants learned the respective linear association from the assigned initial sample: positive initial, b = 2.20, 95% CI [1.68, 2.74], $p < .001$; negative initial $b = -1.51$, 95% CI [−1.72, −1.29], $p < .001$; zero initial $b = .45$, 95% CI [−0.06, 0.95], $p < .001$. This main effect indicated that the observed relationship in Experiment 1 was caused by the statistics hidden in the manipulated initial samples rather than some metapriors. In other words, participants indeed learned the association between an arbitrary facial dimension and behavior from a small sample.

Replicating the sampling interaction in Experiment 1, data from the testing phase found participants in the free sampling condition were more likely to rely on the initial patterns for their final decisions than participants in the fixed sampling condition (Figures 2f, 2g, 2h). In both the positive initial condition (interaction $b = .33$, 95% CI [0.01, 0.64], $p = .043$), and the negative initial condition (interaction $b = -1.01$, 95% CI [−1.33, −0.69], $p < .001$), the interaction effects were statistically significant. Although we expected to see lower variations in the zero initial condition, we observed a significant interaction $b = -.45$, 95% CI [−0.15, −0.75], $p = .003$. Zero initial participants in the fixed sampling condition revealed a significantly negative linear association compared to participants in the free sampling condition, as if the more faces they saw, the more structure they imagined. Last, free-sampling participants viewed on average 11 out of 20 additional samples (Figure 3b).

In sum, we found that participants were able to learn statistical patterns from a small sample of faces and behaviors. The relationship learned can be positive, negative, or zero, which is a function of the statistical patterns encoded into the initial sample. Participants were less likely to revise their previously learned beliefs if they were free to sample faces and behaviors. The lack of sufficient observations impeded accurate learning. In fact, when participants were exposed to the full data set, they were more likely to revise their priors to approach the ground truth, although not perfectly.

Figure 3

The Number of Examples Being Sampled in Free Versus Fixed Sampling Conditions in Experiments 1 and 2



Note. Sample counts in Experiments 1 and 2. The vertical axis indicates the number (mean and 95% confidence intervals) of examples participants viewed in free (red) versus fixed (black) sampling conditions, visually grouped by Experiments 1 and 2 with positive, negative, and zero initial correlations. Dashed lines indicate all samples participants could have seen. As shown, participants in the free sampling condition consistently viewed fewer examples than those in the fixed sampling condition. Pos = positive; Neg = negative. See the online article for the color version of this figure.

Experiment 3: A More Naturalistic Test

The third study aimed to replicate the main results of the first two experiments, but mimicking real-world decisions in a more naturalistic way. We used a subtle manipulation in this study to overcome two concerns. First, in most real-time decisions, people are not typically asked to reveal what they have learned before they have finished learning new things. Moreover, making evaluations before subsequent learning can cause an anchoring effect (Tversky & Kahneman, 1974) that biases the results toward initial pattern confirmation. We, therefore, removed the manipulation check phase. Second, in most real-time decisions, people are rarely forced to view all samples before making decisions (Klein & O'Brien, 2018). We therefore removed the fixed sampling manipulation to reduce demand characteristics. We hypothesized that participants in this study should behave like those in the free sampling condition of Experiments 1 and 2—that they would learn spurious associations, sample fewer faces, and be less likely to revise their impressions. The hypothesis would fail if we did not observe linearity in participants' final decisions, especially for those who are assigned to positive or negative initial sample conditions. We also explored whether individual differences in tolerance toward ambiguity (need for cognitive closure, 15-item short scale; Roets & Van Hiel, 2011) and other social demographics relate to participants' decisions.

Method

Participants

We collected data from 300 new participants (9,000 nested trials) via Amazon Mechanical Turk Prime Cloud Research (i.e., a platform with higher quality participants) for this study with the same power considerations as in the previous two studies. Our participants were 56% male, 73% White, with an average age of 40 and a *SD* of 13, and 47% having earned a bachelor's degree. Free response comments again suggested satisfaction and engagement.

Procedure

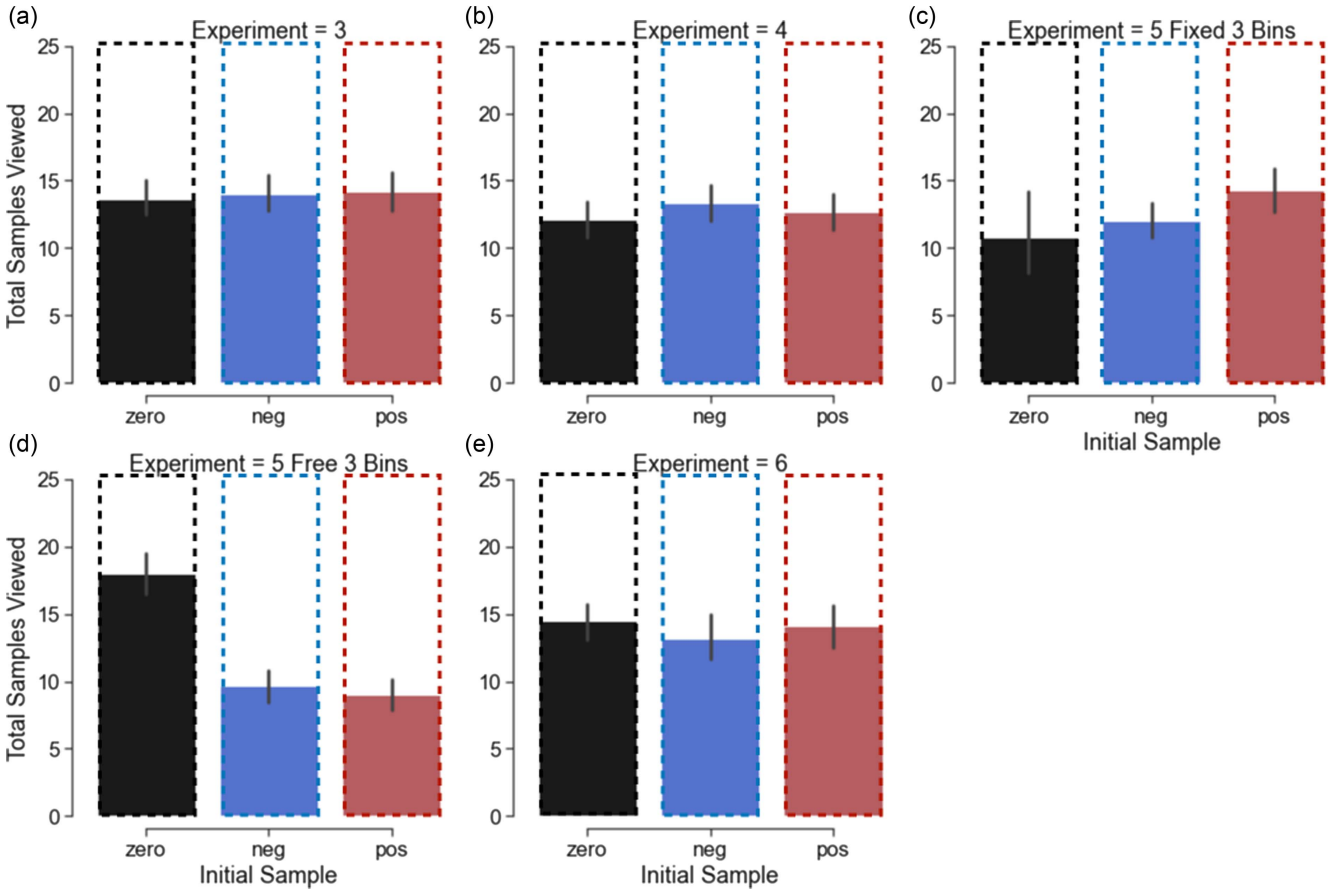
As mentioned, the procedures were identical to Experiment 2, except that we removed manipulation check tests and the fixed sampling conditions. This task took less than 10 min. Participants were introduced to the estimation task of generosity, as in Experiment 2. The instructions stated that they could see up to 25 pictures of attendees and their donations. They saw five faces in the beginning, and then they could continue viewing more faces or stop at any time. Afterward, they made decisions about 30 new attendees. After reading the instructions, participants were randomly assigned to positive ($r = 1.0$, $N = 95$), negative ($r = -1.0$, $N = 100$), and near zero ($r = 0.1$, $N = 105$) initial sample conditions. Right after viewing these initial samples, participants were prompted to freely sample, such that they can "press V" to view more or S to skip to the decision page. Participants then provided their decisions on test faces and completed questionnaires and standard demographic questions.

Results

First, as expected, we replicated the free sampling condition in Experiment 2 (Figure 4a). Data from the testing phase showed that participants who were assigned to the positive initial condition judged new faces based on a positive linear association, $b = 1.33$, 95% CI [1.39, 1.72], $p < .001$. Participants who were assigned to the negative initial condition judged new faces based on a negative linear association, $b = -.54$, 95% CI [−0.70, −0.38], $p < .001$. Participants who were assigned to the near zero initial condition judged new faces based on slightly positive but closer to zero linear

Figure 4

Effects of an Initial Sample Statistic on Final Decisions in Experiments 3–6, That Is, Free Sampling



Note. The critical interaction effects of the initial sample on participants' final decisions from Experiments 3, 4, 5, and 6. The horizontal axis indicates the variation of the test faces, ranging from −8 *SD* to +8 *SD*. The vertical axis indicates estimated amounts of charitable giving for the faces, ranging from $0 to $100. Figures are grouped by Experiment 3 (a), Experiment 4 (b), Experiment 5 with fixed three initial examples (c), Experiment 5 with participant-generated samples (d), and Experiment 6 when participants requested fewer hints (e). Each figure plotted positive (red dot), negative (blue triangle), and zero (black cross) initial conditions with fitted linear regression lines displaying central tendency and 95% confidence intervals. Spurious associations emerged from small samples in most variant conditions, except in Experiment 5 that shows more complicated effects. See the online article for the color version of this figure.

association, $b = .45$, 95% CI [0.07, 0.83], $p < .001$. In addition to the initial five faces, participants on average viewed nine more faces (out of 20) before making decisions (Figure 5a).

We explored which individual-level characteristics correlate with levels of persisting on the initial spurious patterns. We estimated whether a participant persisted in using spurious associations by obtaining their decision responses with 95% confidence intervals. If their decision intervals included 0, this was coded as "no," indicating their decisions did not systematically differ from zero and, thus, were more in line with the population statistics. If their decision intervals did not include 0, this was coded as "yes," indicating their decisions systematically deviated from zero and, thus, were more in line with the initial sample. We ran logistic regressions with various individual predictors. Although the sign of the need for cognitive closure indicated that the more participants were intolerant toward ambiguity, the more likely they were to rely on prior knowledge of some linear relationship, the magnitude was not statistically significant ($b = -.25$, 95% CI [−0.55, 0.05], $p = .108$). Neither length of task time, total number of samples viewed, age, race, gender, nor education level predicted persistence (more details in Supplemental Materials).

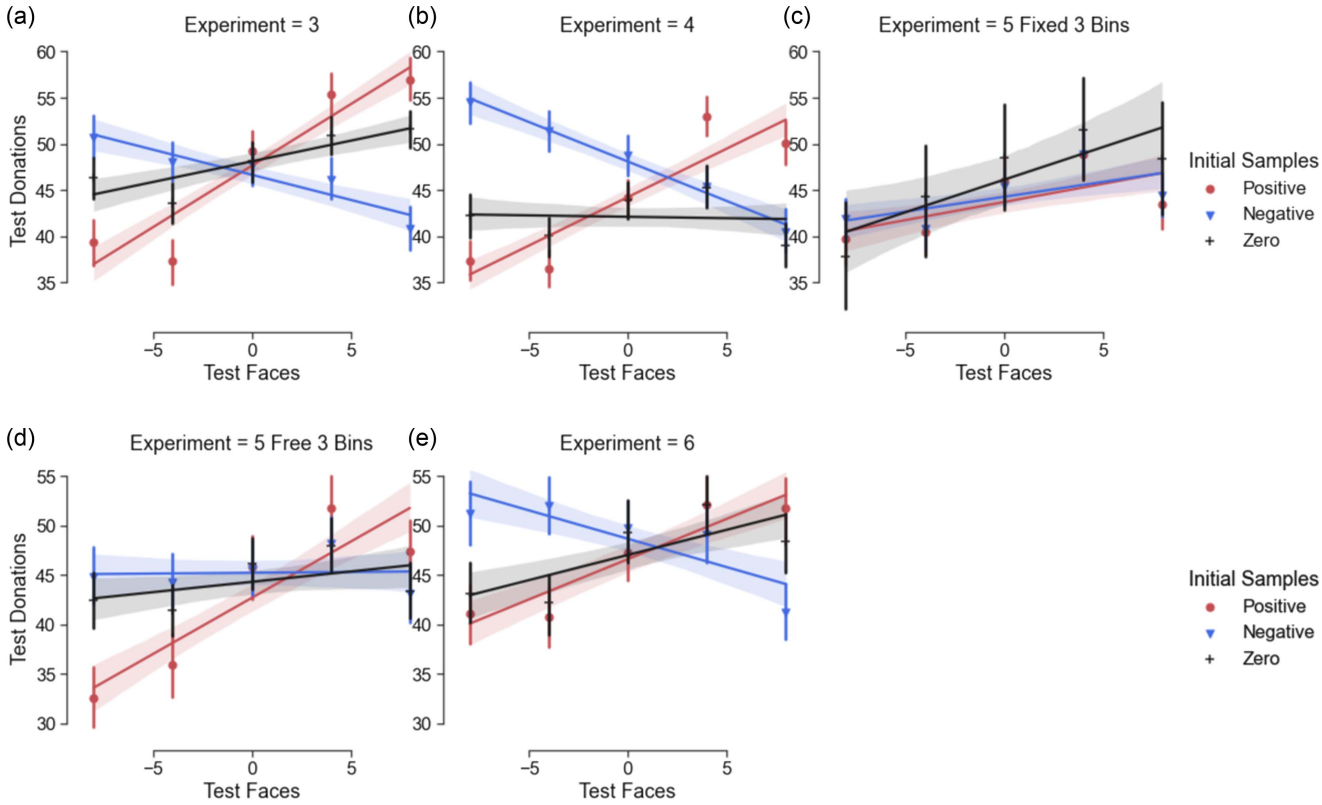
In sum, we replicated the previous findings, using a more naturalistic and subtle design. Participants learned the statistical pattern between novel faces and social behaviors from very few samples—five examples. They were less likely to sample the complete data set, even though they knew it contained only 25 examples. As a result, their final decisions were heavily influenced by the initial samples.

Experiment 4: A Tighter Replication

The fourth study aimed to extend Experiment 3, using a more stringent and less artificial design. First, we added stronger monetary incentives for accuracy. Although we emphasized accuracy throughout the previous studies (e.g., hypothetical trust game earnings in Experiment 1; boldfaced text reminding participants that decisions based on small samples can be misleading), this emphasis was more

Figure 5

The Number of Examples Being Sampled in Experiments 3–6, That Is, Free Sampling



Note. Sample counts in Experiments 3, 4, 5, and 6. The vertical axis indicates the number (mean and 95% confidence intervals) of examples participants viewed in positive (red), negative (blue), and zero (black) initial associations, visually grouped by Experiments 3 (a), 4 (b), 5 with fixed three initial examples (c), 5 with participant-generated samples (d), and additional hint conditions among participants who asked for fewer hints (e). Dashed lines indicate all samples participants could have seen. Overall, participants across all experiments sampled fewer examples than the whole data set, regardless of various incentives and manipulations. Pos = positive; Neg = negative. See the online article for the color version of this figure.

instructional than substantial. Moreover, given that online workers are paid for their time, it is possible that their behaviors (e.g., sampled fewer faces) were driven by higher incentives for speed. To compensate, we added stronger incentives for accuracy with real money.

Second, we revised the instructions to make the task less contrived. Instructions in previous studies and the sorted order of the presentation of the initial stimuli might have missignaled that the initial samples were drawn randomly. These design artifacts could give the impression that the relationship in the initial sample was more or less representative of the population (Tversky & Kahneman, 1974). We, therefore, simplified the instructions and changed stimuli presentations to minimize these issues. The new stimuli were presented in random order.

Third, we reduced the number of initial samples to three. The fact that participants in previous studies learned the hidden statistical relationships from as few as five examples was already impressive, but it might be rare for people to encounter five examples with perfect linearity in rapid succession in real life. Three examples were more plausible than five. We did not know if such a small number of examples is sufficient for people to form any meaningful impressions, so we tested this here. In this study, we also explored how categorical information influences sampling decisions (briefly, we found categorical information exacerbated the effects as predicted; see details in the Supplemental Materials).

We hypothesized that participants in this study should replicate the behaviors observed in Experiment 3 (and in Experiments 1 and 2, in the free sampling condition)—that they would learn spurious associations and sample fewer faces, and as a result would not revise their learned inaccurate impressions. The hypothesis would fail if we did not observe linearity in participants' decisions who were assigned to see the initial three examples containing positive or negative linear associations between faces and behaviors.

Method

Participants

We collected data from 320 new participants (9,600 nested trials) via Amazon Mechanical Turk Prime Cloud Research for this study with the same power considerations as in the previous studies. Our participants were 52% male, 75% White, with an average age of 41 and a *SD* of 13, and 66% having earned a bachelor's degree or higher. Free comments again suggested satisfaction. This task took less than 10 min.

Procedure

As mentioned, the procedures were identical to Experiment 3 with the following modifications: First, before entering the game, participants now saw a bonus instruction page: "Try your best to make accurate decisions. For each correct guess, you earn a $0.05 extra bonus." The research team granted corresponding amounts of bonuses to participants afterward. Note that if participants estimated correctly for all trials, they would have increased their total payoffs by 150%, which makes this accuracy incentive nontrivial. Second, participants now learned that there would be 25 pictures (same as before), and they would first see three selected (instead of "five" or "randomly" selected) pictures. Participants were then randomly assigned to positive, negative, and near zero initial sample conditions, where the three faces were presented in a random order (instead of an ascending order of donation size).

Results

Even after adopting these stringent modifications, participants learned spurious associations and persisted with them in their decisions (Figure 4b). Data from the testing phase showed that participants who were assigned to the positive initial condition ($N = 55$) judged new faces based on a positive linear association, $b = 1.19$, 95% CI [0.97, 1.42], $p < .001$. Participants who were assigned to the negative initial condition ($N = 54$) judged new faces based on a negative linear association, $b = -0.46$, 95% CI [−0.69, −0.24], $p < .001$. Participants who were assigned to the near zero initial condition ($N = 51$) judged new faces based on a slightly positive but closer to zero linear association, $b = .28$, 95% CI [0.04, 0.51], $p = .02$. In addition to the initial three faces, participants on average viewed 10 more faces (out of 22) before making decisions (Figure 5b).

In sum, participants learned spurious associations fast—from as few as three examples. They sampled fewer new faces even when they were incentivized to value accuracy rather than speed. As a result of this fast learning and insufficient sampling, participants persisted in using spurious associations that were consistent with the initial small samples but inconsistent with the larger data set. This finding supports the Insta-learn hypothesis that false face stereotypes are locally accurate but globally inaccurate associations, which can originate from insufficient statistical sampling.

Experiment 5: Idiosyncratic Initial Samples

What do people do when the initial evidence is weak and heterogeneous? So far, we manipulated the initial sample statistics to be either a perfect positive or a perfect negative association. However, people may not always encounter examples with such strong signals for only three identical targets (except manipulated media exposure). In some cases, the strength of the statistical pattern may vary from person to person; the identity being encountered may also vary from person to person. To introduce such variation, this study changed the group-level treatment to an individual-level treatment. Specifically, although each participant saw three initial examples as in the previous study, the three examples were randomly chosen from the entire data set. This change made both the strength of the statistical pattern and the identity of the initial examples variable, making the statistical patterns in the initial sample less obvious and more heterogeneous than in previous experiments.

We hypothesized that participants would still be more likely to use spurious associations and to sample fewer examples. We speculated that participants who have encountered weaker signals in the initial three examples might continue to sample more. If so, we should observe their final decisions to correlate more with their own sampled statistics rather than the three initial examples or the whole population data set. The hypothesis would fail if we observe participants sample almost all examples from the whole data set or if the associations in final decisions do not correlate with any patterns from the initial samples.

Method

Participants

We collected data from 201 new participants (6,030 nested trials) via Amazon Mechanical Turk Prime Cloud Research for this study. Power simulation used empirical results from 31 pilot trials, which suggested a multilevel model of 150 samples should give us a .81 effect size at a .05 α level. Our participants were 51% male, 75% White, with an average age of 41 and a *SD* of 14, and 57% having earned a bachelor's degree or higher. Free comments again suggested satisfaction and instruction clarity. In contrast to the previous studies, some participants indicated that this task was hard. This task took less than 10 min.

Procedure

The procedures and instructions were largely identical to Experiment 4, with one modification in the initial sampling phase: Participants saw three initial pairs of examples, which were randomly drawn from the whole data set of 25 total pairs (instead of fixed pairs). Note that this manipulation yielded a total of 13,800 possible treatments (3 out of 25 possible pairs), and we did not impose any constraints on which three pairs would be shown.

Results

First, consistent with previous experiments, participants on average viewed fewer examples than the total set before making decisions: 13 out of 25 examples in this study (Figure 5c). Second, not many participants gave guesses consistent with the zero correlation in the population in their final decisions. Roughly and descriptively, we found 36 out of 201 participants revealed a relationship between −0.1 and +0.1 as their final decisions, meaning 72% of participants superimposed some linearity in their final decisions.

More precisely, a group-level analysis categorized the statistical pattern in the initial three examples into three bins (as preregistered): a Pearson correlation between faces and donations in the intervals (−1, −0.1), (−0.1, +0.1), (+0.1, +1) was coded as negative ($N = 103$), zero ($N = 17$), and positive ($N = 81$), respectively. We found participants in all three conditions tended to judge new faces based on a positive linear association, $b = .32$, 95% CI [0.15, 0.49], $p < .001$ (Figure 4c). We did not find any statistically significant interactions. In other words, regardless of what participants saw in their initial three examples (positive, negative, or zero with varying strength), participants judged a narrower face to be more generous. We replicated this finding with five fine-grained categories (see preregistration and Supplemental Materials for details). Hence, the positive initial

condition supported our group-level hypothesis, but the negative or the zero initial conditions did not.

In addition to the fixed three initial samples, we explored how participant-generated samples influence their decisions (as preregistered). Each participant chose the number of samples they wanted to see before moving to the decision phase. We calculated a Pearson correlation between faces and donations for each participant using their own samples and categorized the correlation into negative ($N = 68$), zero ($N = 81$), and positive ($N = 52$) conditions. Participants in the self-generated positive condition judged new faces based on a positive linear association, $b = 1.13$, 95% CI [0.89, 1.38], $p < .001$. This was significantly different from participants in the self-generated negative (interaction $b = -1.12$, 95% CI [−1.44, −0.80], $p < .001$) and zero (interaction $b = -.93$, 95% CI [−1.23, −0.62], $p < .001$) conditions. Similarly, we replicated this finding with five fine-grained categories (see Supplemental Materials for details). Hence, participants' decisions in the self-generated positive and zero conditions were consistent with our hypothesis, but participants' behaviors in the self-generated negative condition were inconsistent (Figure 4d).

Although categorizing continuous initial slopes into categories mirrors the analysis of prior experiments, it nonetheless misses granular patterns given that the raw data were on a continuous scale. To test what factors predict participants' decision slopes, for each participant, we calculated their initial slope in the first three examples, the self-generated slope in their sampled examples, and the size of their self-generated samples. We ran a multilevel regression model with the donation decisions as the outcome variable, faces in the decision phase as the predictor variable, and initial slope, self-generated slope, and sample size as additional predictors. Results show that the self-generated slope (interaction term $b = 1.07$, 95% CI [0.691, 1.450], $p < .001$) is a significant predictor of the decision slope, in contrast to the initial slope ($b = 0.131$, 95% CI [−0.055, 0.316], $p = .167$) or the sample size ($b = -0.013$, 95% CI [−0.030, 0.004], $p = .129$). This result indicates that when the initial statistical regularities are weak and heterogeneous, participants may have relied on their own sample statistics to make a decision, although still a biased decision at the individual level.

In sum, we partially replicated our main findings with individual heterogeneous treatments. Each participant saw different pairs of initial samples, randomly drawn from the data set. This manipulation tested how the strength of the initial statistical patterns influences final decisions. With varying identities and varying strengths of the statistical patterns across participants, we confirmed insufficient sampling among participants. Yet, inconsistent with previous experiments, we found participants' final decisions were not related to the initial three examples; they were related to their self-generated samples. One key difference between this experiment and previous experiments is that the initial relationship was much weaker (the average association among the three examples in the positive initial condition was $r = 0.6$ vs. $r = 1.0$ in previous experiments; in the negative initial condition, it was $r = r = -0.7$ vs. -1.0) simply by virtue of the design. The current findings suggest that when the initial relationship was weak, people were willing to sample more, 13 out of 25 examples. Nonetheless, participants still sampled insufficiently, and as a result, their final decisions were related to the patterns from the samples that they themselves generated. In other words, participants may have settled on an association from their own samples, stopped sampling early, and thus persisted in using that spurious association they learned earlier.

Experiment 6: Additional and Reliable Information

What do people do when the actual relationship is zero? Throughout our experiments, we designed the population-level statistics to be zero. However, it is not obvious what people should do in such a situation. If there is no other information available, it is possible that people assume that there must be some relationship and that they settle on the first relationship (positive or negative) they observe in their self-generated samples. This could explain the persistent reliance on spurious associations in our previous experiments. However, if people are provided with additional and reliable information, they need not rely solely on the information gained from observing the face and behavior pairs. To the extent that people discover that the facial cue is useless (e.g., a zero association), they can completely disregard this information and instead rely on additional, reliable information. Hence, this final study aimed to investigate to what extent people utilize additional and reliable information.

Specifically, we provided an opportunity to seek more information during the test phase. Before making each decision, participants were asked whether they wanted to see more information. If they opted to see it, they were provided with information about what the person donated the last time. Unbeknownst to participants, this information was perfectly predictive of what the person donated. Across trials, this additional information also contained the ground truth value of a zero correlation between facial appearance and behaviors at the population level.

We made the following predictions: First, participants in the zero initial condition would be more likely to request more information than participants in the positive or negative linearity conditions because of uncertainty. Second, participants who requested more information would use the additional information to make decisions. As a result, participants in the zero initial condition who asked for more additional information would be less likely to reveal spurious linear associations, as they had access to more information. Third, we predict a main effect between requesting more versus less information. The decision slopes of participants who requested more information should be closer to the ground truth than participants who requested less information across positive, negative, and zero initial conditions. The hypotheses would fail if we did not observe differences in the number of information requests, correlations between additional information and actual decisions, or decision behaviors between the initial zero, initial positive, or initial negative conditions, particularly among participants who requested less additional information than participants who requested more.

Method

Participants

We collected data from 304 new participants (9,120 nested trials) via Amazon Mechanical Turk Prime Cloud Research for this study with the same power considerations as in the previous studies. Our participants were 46% male, 71% White, with an average age of 40 and a *SD* of 14, and 60% having earned a bachelor's degree or higher. Free comments again suggested engagement and attention. This task took about 10 min.

Procedure

As described, this study closely replicated Experiment 4, but with one change in the test phase. To set up the background, in the testing instructions, participants read:

> Hint: For each correct guess, you earn 5 cents extra bonus. Your choices will be counted as accurate with a margin of ±$5. You can also choose to pay 2 cents to see extra information. The extra information is the charitable behavior of the person in a past situation. Press any key to begin.

Next, instead of making decisions directly, participants were asked to first decide, "Do you want more information? Y for yes and N for no." If the Y-key was pressed, participants moved on to the decision trial and saw "Extra information: This person donated x% last time." They made decisions on a slider bar from 0 to 100 above the question, "Out of $100, how much do you think this person donated?" If participants pressed the N-key, they did not see the extra information. This was the same as in Experiment 4, where participants made decisions directly. By design, a random guess without requesting any extra information would result in, on average, a 15-cent bonus. In contrast, an informed guess by requesting extra information on all trials would result in, on average, a 90-cent bonus. Thus, relying on additional information should be helpful and in the interest of participants.

Results

First, replicating previous results, participants in the positive initial condition ($N = 110$) judged new faces based on a positive linear association, $b = .69$, 95% CI [0.52, 0.86], $p < .001$. Participants in the negative initial condition ($N = 91$) judged new faces based on a negative linear association, $b = -.41$, 95% CI [−0.60, −0.22], $p < .001$. Participants in the zero initial condition ($N = 103$) judged new faces based on a slightly positive linear relationship, $b = .46$, 95% CI [0.28, 0.63], $p < .001$. Similar to previous experiments, in addition to the initial three faces, participants on average viewed 11 more faces before making decisions (Figure 5e). Next, we examined the utility of the additional information.

Our first hypothesis was not supported (Figure 5e). We hypothesized that participants in the initial zero condition should be more likely to request additional information than participants in the initial positive or negative conditions, but results revealed no differences between conditions (zero initial $b = 9.05$, 95% CI [3.23, 14.86], negative initial $b = 9.56$, 95% CI [7.11, 12.01], positive initial $b = 10.76$, 95% CI [5.00, 16.53], all pairwise comparisons *ns*). Descriptively, about two thirds of the participants in all three conditions requested information fewer than 10 times, and about one third requested more than 20 times (out of 30 total).

Our second hypothesis was supported. As predicted, there was a strong positive linear association between hints and actual decisions: Multilevel regression with hints as the independent variable and actual decision as the dependent variable clustered within each participant revealed a statistically significant correlation: $b = 0.99$, 95% CI [0.97, 1.00], $p < .001$. Regardless of the initial sample statistics, as long as participants requested additional information, they relied heavily on that new information to make decisions.

Our third hypothesis was supported as well. Among participants who searched for more hints (i.e., requested more than 20 hints out of 30 total; preregistered), there were no initial condition differences in their final decisions: initial negative $b = -0.12$, 95% CI [−0.50, 0.26], $p = .55$; which differed, but not significantly, from the initial positive by $b = .44$, 95% CI [−0.07, 0.95], $p = .09$; and also not statistically significantly from the initial zero by $b = .29$, 95% CI [−0.25, 0.83], $p = .30$. In other words, most of them used the information given and ignored the face information, resulting in decisions similar to the population statistics. In contrast, participants who searched for fewer hints (i.e., requested fewer than 10 hints; preregistered) showed systematic persistence of the initial sample statistics in their final decisions (Figure 4e): initial negative $b = -.57$, 95% CI [−0.80, −0.34], $p < .001$; which significantly differed from the initial positive by $b = 1.39$, 95% CI [1.07, 1.70], $p < .001$; and from the initial zero by $b = 1.08$, 95% CI [0.77, 1.40], $p < .001$. In other words, participants who had seen initial examples containing negative/positive/zero linearity without asking for more information persisted in using spurious associations.

In sum, additional reliable information does make a difference, although not for all participants. Participants who asked for more direct information made more accurate predictions than participants who did not ask for more information. Nonetheless, not asking for more information seems to be the default, as indicated by two thirds of our participants, regardless of what they initially saw. Despite its effectiveness, asking for more information seems to be the exception. Fast learning with few initial examples and insufficient sampling in the face of potentially reliable information again generated false face stereotypes.

Moderator Analyses

So far, our data show that participants formed inaccurate impressions when assigned to conditions where the initial small samples contained strong statistical regularities, and they sampled insufficiently. Three factors may affect their decisions: the statistical regularities in the initial sample, the statistical regularities in the samples they generated, and the size of the self-generated samples. This section analyzes which factor(s) contribute more to the statistical patterns observed in participants' final decisions. Given that the setting in Experiment 6 is procedurally different from the other experiments, here we use data from Experiments 1 to 5.

Method

For each experiment within each experimental condition, we ran a multilevel regression model with faces in the decision phase as the predictor variables and donations as the outcome variables (identical to the statistical analysis above). Here, we added the slope in the initial samples, the slope in the self-generated samples, and the sample size as the moderators to the baseline model. The most informative statistical information is the interaction term *b*, which indicates to what extent participants' decision slopes are *more* or *less* influenced by each of the above moderators.

Results

As shown in Table 1, within each experimental condition, there is at least one factor that moderates the effects. First, the initial slope was the most robust and consistent predictor of the decision slope: Except for Experiment 5 (the heterogeneity condition), the initial

Table 1
Initial Slope, Self-Generated Slope, and Sample Size as Predictors of Decision Slope

Experiment	Condition	Initial slope	Self-generated slope	Sample size
1	Free	**b = 0.775, 95% CI [0.664, 0.885], p < .001**	**b = 1.938, 95% CI [1.495, 2.381], p < .001**	b = 0.005, 95% CI [−0.002, 0.011], p = .168
2	Positive	**b = 0.610, 95% CI [0.392, 0.827], p < .001**	**b = 2.349, 95% CI [1.655, 3.043], p < .001**	**b = −0.078, 95% CI [−0.111, −0.045], p < .001**
	Negative	**b = −0.653, 95% CI [−0.865, −0.442], p < .001**	b = 0.138, 95% CI [−0.527, 0.803], p = .684	b = −0.005, 95% CI [−0.036, 0.026], p = .739
	Zero	b = 0.138, 95% CI [−0.073, 0.350], p = .2	b = −0.708, 95% CI [−2.153, 0.738], p = .338	**b = −0.055, 95% CI [−0.082, −0.027], p < .001**
3	Positive	**b = 1.331, 95% CI [1.164, 1.498], p < .001**	**b = 0.855, 95% CI [0.355, 1.356], p < .001**	**b = −0.028, 95% CI [−0.051, −0.005], p = .016**
	Negative	**b = −0.543, 95% CI [−0.702, −0.383], p < .001**	b = 0.067, 95% CI [−0.422, 0.555], p = .789	b = −0.008, 95% CI [−0.014, 0.030], p = .456
	Zero	**b = 0.447, 95% CI [0.298, 0.596], p < .001**	b = 0.930, 95% CI [−0.006, 1.866], p = .052	b = −0.001, 95% CI [−0.023, 0.021], p = .937
4	Positive	**b = 1.049, 95% CI [0.886, 1.212], p < .001**	**b = 1.244, 95% CI [0.781, 1.707], p < .001**	**b = −0.078, 95% CI [−0.099, −0.056], p < .001**
	Negative	**b = −0.854, 95% CI [−1.017, −0.692], p < .001**	**b = 1.639, 95% CI [1.168, 2.110], p < .001**	**b = 0.038, 95% CI [0.018, 0.059], p < .001**
	Zero	b = −0.028, 95% CI [−0.192, 0.137], p = .235	**b = 1.324, 95% CI [0.569, 2.079], p < .001**	b = 0.009, 95% CI [−0.016, 0.033], p = .489
5	All	b = 0.131, 95% CI [−0.055, 0.316], p = .167	**b = 1.070, 95% CI [0.691, 1.450], p < .001**	b = −0.013, 95% CI [−0.030, 0.004], p = .129

Note. The interaction term, b, between sampling condition and initial slope, self-generated slope, and sample size, reported with point estimates and 95% confidence intervals and p values. Values in bold indicate significance at p < .05. CI = confidence interval.

slope statistically significantly predicted participants' donations (see column "Initial Slope").

Second, the self-generated slope consistently moderated the decision slope when the initial sample had five and 11 examples showing a positive, but not negative, or zero, slope (see Experiments 1–3, positive conditions). The self-generated slope played a more important role when the initial sample had three examples or when it did not contain strong statistical patterns (see Experiments 4 and 5, all conditions).

Third, the size of the self-generated sample showed less consistent patterns: When the slope in the initial sample was positive, the more examples participants saw, the less likely they were to use positive slopes in their decisions. However, similar effects were not obtained in the negative or zero conditions.

In short, this analysis suggests that all three factors—initial slope, self-generated slope, and sample size—can contribute to participants' final decisions that show spurious correlations. However, the statistical regularities in the initial sample are likely the most consistent moderators in the tested studies.

Simulated Benchmarks

A critical assumption of our experiments is that participants sample insufficiently. However, given the relatively small number of face behavior pairs and the strong initial statistical patterns, it may be that this sampling behavior is optimal. In this section, we formulate five benchmarks simulating optimal behaviors in equivalent settings and compare these simulated behaviors against our empirical data.

Method

Overall, we simulated two kinds of benchmarks, with one reflecting the experimenter's perspective and the other reflecting the participant's perspective. In both sets of simulations, the decision-maker uses online Bayesian linear regression to estimate the expectations of the coefficients between facial images and donation behaviors. On the one hand, the experimenter knows the ground truth; therefore, the optimal behavior from the lens of the experimenter is to stop sampling when the decision-maker accurately learns the statistical pattern. Accuracy is the goal here. We simulate three scenarios from more conservative to more liberal: (a) allowing no errors in estimations, (b) adding error bounds, and (c) adding sampling costs. On the other hand, the participant does not know the ground truth; therefore, the optimal behavior from the lens of the participant is to stop sampling when the decision-maker feels they have learned enough. We formalize this experience as changes in the expected coefficients after integrating new samples. When the change is smaller than a certain threshold, we can benchmark that the decision-maker has learned enough and, thus,

it is optimal to stop. Efficiency is the goal here. We simulate two scenarios: (d) setting thresholds for changes and (e) adding sampling costs.

Results

We first use data from Experiment 3, positive initial correlation, to highlight the key takeaways while demonstrating the analysis details. We then present a summary table of other experiments in Table 2, showing the optimal size and corresponding coefficients, as well as the empirical sample size and the empirical coefficients.

Overall, as compared to the optimal benchmarks, participants overestimated the coefficients, thus they learned too much (Benchmarks No. 1, No. 2, No. 3, No. 4, No. 5). Without considering the costs for sampling, participants sampled fewer than the optimal size, thus they sampled too little (Benchmarks No. 1, No. 2, No. 4). Including the costs for sampling, under certain conditions (i.e., unit cost, cost function, accuracy-cost trade-off function, error bounds, and delta thresholds), participants might have sampled just enough (Simulations No. 3, No. 5).

Benchmark No. 1

Allowing no errors. Figure 6a simulates that the more examples the decision-maker samples (on the horizontal axis), the closer their expected coefficient is to being zero (on the vertical axis). The red lines represent the "optimal" choice, which is to sample $n' = 20$ more examples to achieve an accurate estimation of $\beta' = 0$. In comparison, the green lines represent the "empirical" choice, which is the average responses from our participants in Experiment 3, the positive initial condition. Empirically, our participants sampled $n = 9$ more examples, estimated $\beta = 0.98$ in the sampling phase, and estimated $\beta = 1.33$ in the testing phase. Both estimations are 1 *SD* away from the ground truth. This simulation suggests that human participants might have learned too much ($\beta = 1.33$ vs. $\beta' = 0$) while sampling too little ($n = 9$ vs. $n' = 20$).

Benchmark No. 2

Adding error bounds. However, some may argue that allowing no errors is too stringent. Therefore, Figure 6b plots the number of optimal samples, allowing estimation errors to occur within a certain bound. There are many ways to define an optimal bound. Here, we use responses in the zero initial condition as the optimal comparison. Participants in the zero initial condition estimated the coefficient as 0.45. As compared to the ground truth coefficient of 0, the error $epsilon' = 0.45$. Now, we can ask: How many samples are needed to achieve an error that is equal to or smaller than the optimal bound? The simulation shows that it is needed to sample at least $n' = 15$ examples to achieve such bound (red lines). However, participants empirically sampled $n = 9$ and the error $epsilon = 1.33$ (green lines), leading to a larger error bound than the optimal bound. This simulation suggests that even allowing comparable errors as in the zero initial condition, human participants might have learned too much ($epsilon = 1.33$ vs. $epsilon' = 0.45$) while sampling too little ($n = 9$ vs. $n' = 15$).

Using Benchmarks 1 and 2, we can calculate the optimal sample size and estimated coefficient for each condition from Experiments 1 to 4 and compare the optimal values against the empirical data. Given that this is a group-level analysis, we have yet to consider the heterogeneous priors in the simulation. In this table, we only used Benchmarks 1 and 2 because the assumptions for the parameters are

Table 2
Comparing Simulated and Empirical Sizes and Coefficients

Experiment	Condition	Benchmark	Optimal n'	Empirical n	Optimal b'	Empirical b
1	Free	No. 1	110	25	0	0.775
2	Positive	No. 1	20	6	0	0.610
	Negative	No. 1	20	5	0	−0.653
	Zero	No. 1	20	6	0	0.138
3	Positive	No. 1	20	9	0	1.331
	Negative	No. 1	20	9	0	−0.543
	Zero	No. 1	20	9	0	0.447
4	Positive	No. 1	22	10	0	1.049
	Negative	No. 1	22	11	0	−0.854
	Zero	No. 1	22	9	0	−0.030
1	Free	No. 2	66	25	0.303	0.775
2	Positive	No. 2	18	6	0.138	0.610
	Negative	No. 2	18	5	−0.138	−0.653
	Zero	No. 2	20	6	0	0.138
3	Positive	No. 2	15	9	0.447	1.331
	Negative	No. 2	15	9	−0.447	−0.543
	Zero	No. 2	20	9	0	0.447
4	Positive	No. 2	21	10	0.030	1.049
	Negative	No. 2	21	11	−0.030	−0.854
	Zero	No. 2	20	9	0	−0.030

Note. Summary results are shown in this table. See example walk-throughs in the main text and analysis code to reproduce the simulation numbers in the online repository (https://osf.io/syc6b/?view_only=555262392d9b4943b2145c0ed00efd23).

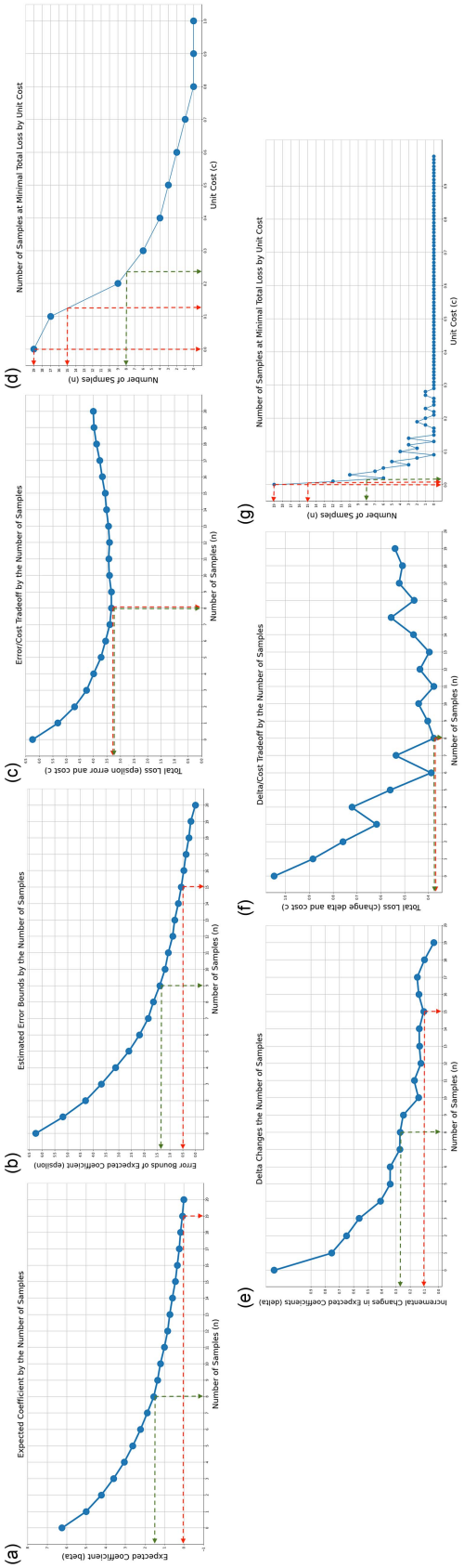


Figure 6
Simulated Benchmarks From Accuracy-Driven Experimenters' Perspective (a–d) and Efficiency-Driven Participants' Perspective (d–g)

Note. See detailed explanations in the main text. See the online article for the color version of this figure.

minimal. More generalized patterns in Benchmarks 3–5, which require more subjective assumptions, are presented in the figures below.

Benchmark No. 3

Adding costs. The previous simulations did not consider sampling costs. Although not explicitly penalized, it is plausible that participants might have incurred some costs in drawing new examples. Therefore, Figures 6c and 6d simulate the optimal number of samples as a function of the errors in the expected coefficients and the psychological costs. Given that costs increase linearly with sample size whereas errors decrease exponentially, we can construct the decision as the linear combination between accuracy and cost and calculate the optimal size. Figure 6c shows a specific case for the optimal sample size when the unit cost equals 0.2. Under this cost term and the linear combination assumption, we observe that the optimal sample size n' converges to the empirical sample size $n = 9$, with an estimated optimal coefficient of $\beta' = 1.56$, as compared to the empirical $\beta = 1.33$. As indicated by the overlapping green and red lines. Note that the unit value of the cost term depends on a set of assumptions, including the form of the cost term, the linear combination assumption, and any omitted variables. Therefore, in this simulation we only use the cost value for illustration purposes; this is not comparable to the incentive designed in the experiment.

More generally, Figure 6c shows the optimal sample size as a function of varying levels of costs. As shown, the higher the unit cost (horizontal axis), the smaller the optimal sample size is (vertical axis). Again, green lines represent the empirical observations, whereas red lines represent the optimal sample size when cost equals zero (Figure 6a) and when the error bound is within the zero initial condition (Figure 6b). Under very specific conditions such as Figure 6d, one may argue that participants sampled just enough (vs. too little, $n = n' = 9$), but the consequence is still learning too much (vs. accurately $\beta = 1.33$ vs. $\beta' = 0$).

In sum, the psychological costs of sampling dynamically change what an optimal sample size would be, depending on the unit cost and the specific form of the loss function. Regardless of whether the sample size is optimal or not, participants do not learn accurately, enabling inaccurate face impressions to persist in decisions.

Benchmark No. 4

Setting delta thresholds. Delta is defined as the incremental change between the estimated coefficient at time $t - 1$ and the estimated coefficient at time t. Figure 6e simulates when the decision-maker slows down updating their beliefs about the coefficient. For example, under the condition when the delta threshold is set to be smaller than 0.1, the red line represents an example slowdown: Since after the decision-maker sampled $n' = 15$ examples, the changes in the estimated coefficients become smaller, $\delta' = 0.09$. In comparison, our participants empirically sampled $n = 9$ examples with a wider room for changes, $\delta = 0.27$. Generally speaking, the number of optimal samples changes dynamically as a function of how the decision-maker subjectively defines their delta thresholds. This simulation suggests that, depending on the delta thresholds participants have in their mind, some of them may have sampled too little ($n = 9$ vs. $n' = 15$) and learned too much ($\delta = 0.27$ vs. $\delta' = 0.09$).

Benchmark No. 5

Adding sampling costs. We can add cost terms to the above simulation. Figure 6f characterizes a decision-maker who tracks if their updates in the expected coefficients are large enough, that is, the delta, and if adding more samples, that is, the cost, might outweigh the delta benefits. As shown in Figure 6f, with a unit cost of 0.0025 (this is because the delta range is much smaller), we observe it is optimal to sample $n' = 9$ examples, which is the same as the empirical data sampled $n = 9$ (the overlapping lines). Under this set of assumptions (unit cost, quadratic terms of the cost, a linear combination of delta change and cost, subjective beliefs on delta change, etc.), one may argue that participants indeed sampled just enough. Nonetheless, participants were unable to recover the true coefficient due to early stopping. Finally, Figure 6g shows that generally, as the cost increases, the optimal number of samples decreases, reflecting the accuracy-cost trade-off.

Taken together the five benchmarks, three from the experimenter's perspective with the goal of learning accurately (i.e., minimizing the difference between the estimated coefficient and the true coefficient) and two from the decision-maker perspective with the goal of learning efficiently (i.e., minimizing the difference between the estimated coefficient from the last and current pairs of samples), we offer three takeaways: First, as compared to the benchmarks, participants overestimated the coefficients, and thus they learned too much (Benchmarks No. 1, No. 2, No. 3, No. 4, No. 5). Second, as compared to the benchmarks, without the costs for sampling, participants sampled fewer than the optimal size, and thus they sampled too little (No. 1, No. 2, No. 4). Third, as compared to the benchmarks, including the costs for sampling, under certain conditions (i.e., unit cost, cost function, accuracy-cost trade-off function, error bounds, and delta thresholds), participants might have sampled just enough (No. 3, No. 5).

General Discussion

Six experiments consistently supported the Insta-learn account as one potential mechanism leading to the formation and persistence of false face stereotypes. First, participants learned quickly, extracting statistical patterns between arbitrary facial appearances and social behaviors from as few as three examples. Second, participants did not sample sufficiently when given the opportunity to freely decide how many more examples they wanted to see. Due to the combination of fast statistical learning and insufficient sampling, participants in the free sampling condition failed to revise the previously learned spurious associations. This effect was robust across studies and persisted even when participants were incentivized for accuracy. When the initial statistical patterns were weaker, participants went beyond the initial three examples but still did not reach sufficiently large samples, generating self-induced spurious correlations. When provided with additional direct and reliable information, participants were able to revise their prior spurious impressions as long as they asked for that information. However, most participants rarely asked for that information, and, as a result, their spurious impressions persisted. The findings were not moderated by individual-level characteristics such as age, gender, ethnicity, social class, or the need for cognitive closure.

Insta-learn is a domain-general process, consistent with the overgeneralization account, such that the initial sample from a particular population contains certain statistical relationships (e.g., babies and submissiveness). However, the initial sample does not necessarily represent the larger population (e.g., babies are not representative of baby-faced adults). When people do not sample enough (e.g., stop observing or interacting with baby-faced adults), they can continue to rely on spurious impressions (e.g., baby-faced adults are submissive). This mechanism is applicable to many domains, including fitness, familiarity, and emotion overgeneralization, and extends to arbitrary domains just as the examples we used throughout this article. As such, it is not constrained to specific face stereotypes, which might have evolutionary roots. The ability to quickly learn statistical regularities, coupled with insufficient sampling, can make even arbitrary relationships persist. Our experimental design identifies biased initial samples and insufficient sampling (i.e., free vs. fixed sampling) as causal mechanisms.

Of course, we do not think Insta-learn is the only way that face stereotypes emerge and persist. However, it provides one plausible mechanism when there is a mismatch between initial samples and the population. We designed the initial sample to be highly unrepresentative and biased in Experiments 1–4, but this design may reflect real-world cases when there is a shared cultural, though biased, consensus expressed in skewed samples. For example, we are rarely exposed to a random or representative sample of the population, just as babies are not representative of baby-faced adults and family relatives are not representative of familiar-looking people. In fact, unrepresentative, biased, small samples may be a feature of social environments rather than a bug (Fiedler, 2000). In some sense, it is hard to imagine how people would *not* make errors in such a wicked environment (Hogarth et al., 2015). Insta-learn predicts that if the initial samples are representative of the larger population, then people should be less likely to form face stereotypes, just like those participants in our zero initial conditions and some participants in Experiment 5, where the initial samples were highly heterogeneous. The mismatch between the initial sample and the larger population is at the heart of the Insta-learn mechanism. This mismatch can be found in real life. Police Facebook posts in the United States overrepresent Black suspects relative to local arrest rates (Grunwald et al., 2022), and local news in Los Angeles and Orange counties is significantly more likely to portray Blacks and Latinos as lawbreakers than Whites (Dixon, 2006), signaling a strong race-crime association. What our paradigm shows is that when presented with strongly biased initial samples and subsequently more moderate samples, people were not particularly sensitive to these inconsistencies and, consequently, did not update their decisions sufficiently. It is an interesting research question under what conditions people update their beliefs sufficiently.

We also demonstrated a robust effect of insufficient sampling, even with accuracy incentives or access to reliable information. The question "why" people do not sample enough needs more investigation. One plausible mechanism is that people sample very few cases when they see a strong, not weak, statistical relationship in the initial sample. However, this hypothesis is not supported by the current data. We found participants in Experiments 2–4 free sampling conditions sampled similarly, around 8–10 examples, regardless of whether they were assigned to a strong signal (linear

positive or linear negative) or a weak signal (zero) condition. Likewise, in Experiment 5 with varying levels of signals and varying face identities, the correlation between the number of sampled examples and the statistical patterns in the initial sample was essentially zero, $r = 0.04$, $p = .53$. Similarly, the number of sampled examples and the statistical patterns in the self-generated sample were also null, $r = 0.07$, $p = .32$. One may argue that rather than the initial signal, it could be the next sample right after the initial examples that confirms the pattern that motivates stopping. With an extended initial sample, we still do not see a robust relationship: Experiment 5, with 1 more example, $r = 0.03$, $p = .63$; with two more examples, $r = -0.09$, $p = .22$. Our interpretation of these new analyses is that sample size per se may not be an operating variable in our paradigm, but the *statistical regularities* in the initial and/or self-generated samples have lasting effects. Although our current moderator analyses explored this direction and found a unique contribution of the statistical regularities in the initial sample, future work should design careful experiments to tease apart these mechanisms.

Another plausible mechanism includes the attitude asymmetry theory (Fazio et al., 2004), which hypothesizes that people stop sampling to avoid negative outcomes. However, participants in our experiments stopped sampling early in the absence of any feedback. Theories on outcome-dependent sampling suggest that unjustified confidence may result from a lack of disconfirming feedback (Einhorn & Hogarth, 1978). Yet, the question remains as to whether people feel confident even though they have only seen three examples. The motivated tactician theory would suggest that participants' motivation is insufficient, and that is why people do not bother to continue sampling (Fiske & Neuberg, 1990), but monetary incentives were insufficient to induce more sampling. Future work can extend this paradigm and study the optimal incentives that can overcome any psychological costs for sufficient sampling.

Although our studies provide experimental evidence that strong statistical regularities in a very small sample can perpetuate face stereotypes, we do not claim the evidence is ecologically valid. Both internal and external validity are important criteria for establishing a generalizable theory; the present study presents a generalizable and quantitative baseline for future investigation (Banaji & Crowder, 1989). When examining Insta-learn in the field, one critical assumption is that the association between many behaviors and facial appearances is essentially zero. It is under those stringent conditions that we observed participants construct spurious associations due to insufficient sampling. However, some may believe that there are true appearance–characteristic signals. This is ultimately an open and empirical question that needs more research (Todorov, 2017). We emphasize a subtle but important difference between an ideal, nonmanipulated environment (i.e., the world in our study) and an environment that is already tainted by stereotypes (i.e., the world we live in). Consider our experiments in which participants conclude that wider (vs. narrower) faces are more generous after seeing only three examples. If they were to act on these beliefs, they would invite wider faced (and exclude narrower faced adults) to future events. For new participants attending these events, this biased sample would confirm the spurious association. In a way, the choices of the initial participants are "tainting" the environment. Empirical support for this mechanism is demonstrated by iterated learning experiments (Uddenberg et al., 2023). When participants' judgments of associations between faces and behaviors serve as an input to the judgments of other participants, spurious associations reflecting prior biases quickly emerge. That is, even if there are appearance–character associations in the wild, those associations can be the consequences, not the antecedents, of insufficient statistical learning. Other ecological validity limitations include that we only used a specific behavioral tendency and specific variation in facial features. Hence, the generalizability to other behaviors or facial features is not yet established. Using our work as a baseline, future work can test whether valence (positive vs. negative) and particular social dimensions (warmth vs. competence) moderate the observed effects (see Table 3 for more limitations).

Insta-learn examines a boundary condition of statistical learning in traditional cognitive science. The question being asked here is as follows: When there is a mismatch between sample statistics and population statistics, what do humans learn? Humans make sense of the world from limited amounts of data (Tenenbaum et al., 2011). It is an incredible ability but one that can have detrimental consequences under certain circumstances. In classic cognitive tasks, samples are often unbiased and resemble real-world statistics, such as object names or grammar (Saffran et al., 1996; Xu & Tenenbaum, 2007). However, in many social domains, samples do not necessarily resemble real-world statistics. As intelligent humans, people learn the association between particular traits and facial features quickly and accurately. However, as overconfident humans, they stop learning too early. If people are less likely to interact with those groups offline or less likely to search about those groups online just as participants in our Experiment 6, face stereotypes would remain unchallenged. If the sample and the population statistics do not match, the same cognitive ability that makes humans intelligent can also create social problems.

Uncovering the emergence and persistence of face stereotypes from fast statistical learning and insufficient sampling can offer policy implications. Adding to the growing literature on structural changes in diversity science (Banaji et al., 2021; Onyeador et al., 2021; Skinner-Dorkenoo et al., 2023), our work provides causal evidence that changing representations might be effective. Recent analysis shows that increased representational diversity correlates with decreased stereotypes (Bai et al., 2020; Eagly & Koenig, 2021), but we offer more nuanced practices. First, practitioners should think about how to craft a good initial sample: those who appear in public presentations, who interact with clients, who recruit new employees, and so on. The weaker the statistical relationship between demographic features and personal characteristics, the less likely people would be to develop biases. Second, interventions should discuss how to encourage continuous sampling. If not asking for more information is the default behavior, as in Experiments 5 and 6, bias can develop quickly without sustained sampling. Creating institutional rules like our fixed-sampling condition is one idea, but enforcement undermines autonomy (Leslie, 2019). How to overcome insufficient sampling is worth investigating further.

Table 3
Table of Limitations

Dimension	Question	Assessment
Internal validity	Is the phenomenon diagnosed with experimental methods?	Yes.
	Is the phenomenon diagnosed with longitudinal methods?	No.
	Were the manipulations validated with manipulation checks, pretest data, or outcome data?	Yes.
	What possible artifacts were ruled out?	We ruled out the possibility that our results were due to participants' demographic features of age, race, gender, socioeconomic status, and the need for cognitive closure. We also ruled out the societal assumptions associated with any particular facial structure by using nonsocial dimensions as the manipulation.
Statistical validity	Was the statistical power at least 80%?	Yes.
	Was the reliability of the dependent measure established in this publication or elsewhere in the literature?	Yes.
	If covariates are used, have the researchers ensured they are not affected by the experimental manipulation before including them in comparison across experimental groups?	Yes.
	Were the distributional properties of the variables examined and did the variables have sufficient variability to verify effects?	Yes.
Generalizability to different methods	Were different experimental manipulations used?	Yes.
Generalizability to field settings	Was the phenomenon assessed in a field setting?	No.
	Are the methods artificial?	Yes in the sense that this study is done with precisely manipulated variables. No in the sense that this sequence of study relaxes "artificiality" from various aspects.
Generalizability to times and populations	Are the results generalizable to different years and historical periods?	Cannot answer with the current data. This is an open question.
	Are the results generalizable across populations?	Cannot answer with the current data. This is an open question.
Theoretical limitations	What are the main theoretical limitations?	Our argument centers around false face stereotypes; however, one could argue that the same cognitive mechanism applies to other social domains. To what extent this theory applies to other social domains remains an open question.

Conclusion

It is often celebrated that humans learn *so* much from *so* little. It is an incredible ability that allows human societies to thrive. However, maybe humans also learn *too* much from *too* little. The ability to learn quickly from small numbers of examples and insufficiently sample more examples can generate and perpetuate inaccurate face stereotypes.

References

Albohn, D. N., & Adams, R. B., Jr. (2020). Everyday beliefs about emotion perceptually derived from neutral facial appearance. *Frontiers in Psychology*, *11*, Article 264. https://doi.org/10.3389/fpsyg.2020.00264

Albohn, D. N., Uddenberg, S., & Todorov, A. (2022). A data-driven, hyper-realistic method for visualizing individual mental representations of faces. *Frontiers in Psychology*, *13*, Article 997498. https://doi.org/10.3389/fpsyg.2022.997498

Bai, X., Fiske, S. T., & Griffiths, T. L. (2022). Globally inaccurate stereotypes can result from locally adaptive exploration. *Psychological Science*, *33*(5), 671–684. https://doi.org/10.1177/09567976211045929

Bai, X., Ramos, M. R., & Fiske, S. T. (2020). As diversity increases, people paradoxically perceive social groups as more similar. *Proceedings of the National Academy of Sciences*, *117*(23), 12741–12749. https://doi.org/10.1073/pnas.2000333117

Banaji, M. R., & Crowder, R. G. (1989). The bankruptcy of everyday memory. *American Psychologist*, *44*(9), 1185–1193. https://doi.org/10.1037/0003-066X.44.9.1185

Banaji, M. R., Fiske, S. T., & Massey, D. S. (2021). Systemic racism: Individuals and interactions, institutions and society. *Cognitive Research: Principles and Implications*, *6*(1), Article 82. https://doi.org/10.1186/s41235-021-00349-3

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Bjornsdottir, R. T., Hensel, L. B., Zhan, J., Garrod, O. G. B., Schyns, P. G., & Jack, R. E. (2024). Social class perception is driven by stereotype-related facial features. *Journal of Experimental Psychology: General*, *153*(3), 742–753. https://doi.org/10.1037/xge0001519

Bott, F. M., & Meiser, T. (2020). Pseudocontingency inference and choice: The role of information sampling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(9), 1624–1644. https://doi.org/10.1037/xlm0000840

Chua, K. W., & Freeman, J. B. (2022). Learning to judge a book by its cover: Rapid acquisition of facial stereotypes. *Journal of Experimental Social Psychology*, *98*, Article 104225. https://doi.org/10.1016/j.jesp.2021.104225

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. In J. M. Olson (Ed.), *Advances in experimental social psychology* (pp. 131–199). Elsevier Academic Press. https://doi.org/10.1016/bs.aesp.2017.03.001

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review*, *112*(4), 951–978. https://doi.org/10.1037/0033-295X.112.4.951

Dixon, T. L. (2006). Psychological reactions to crime news portrayals of Black criminals: Understanding the moderating roles of prior news viewing and stereotype endorsement. *Communication Monographs*, *73*(2), 162–187. https://doi.org/10.1080/03637750600690643

Dotsch, R., Hassin, R. R., & Todorov, A. (2016). Statistical learning shapes face evaluation. *Nature Human Behaviour*, *1*(1), Article 0001. https://doi.org/10.1038/s41562-016-0001

Eagly, A. H., & Koenig, A. M. (2021). The vicious cycle linking stereotypes and social roles. *Current Directions in Psychological Science*, *30*(4), 343–350. https://doi.org/10.1177/09637214211013775

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, *85*(5), 395–416. https://doi.org/10.1037/0033-295X.85.5.395

Engell, A. D., Todorov, A., & Haxby, J. V. (2010). Common neural mechanisms for the evaluation of facial trustworthiness and emotional expressions as revealed by behavioral adaptation. *Perception*, *39*(7), 931–941. https://doi.org/10.1068/p6633

Evans, N. J., Bennett, A. J., & Brown, S. D. (2019). Optimal or not; depends on the task. *Psychonomic Bulletin & Review*, *26*(3), 1027–1034. https://doi.org/10.3758/s13423-018-1536-4

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology*, *87*(3), 293–311. https://doi.org/10.1037/0022-3514.87.3.293

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*(4), 659–676. https://doi.org/10.1037/0033-295X.107.4.659

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 357–411). McGraw-Hill.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). Academic Press. https://doi.org/10.1016/S0065-2601(08)60317-2

Fiske, S. T., & Taylor, S. E. (1984). *Social cognition* (1st ed.). Mcgraw-Hill Book Company.

Grunwald, B., Nyarko, J., & Rappaport, J. (2022). Police agencies on Facebook overreport on Black suspects. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *119*(45), Article e2203089119. https://doi.org/10.1073/pnas.2203089119

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology*, *12*(4), 392–407. https://doi.org/10.1016/S0022-1031(76)80006-6

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Hill, T., Lewicki, P., Czyzewska, M., & Schuller, G. (1990). The role of learned inferential encoding rules in the perception of faces: Effects of nonconscious self-perpetuation of a bias. *Journal of Experimental Social Psychology*, *26*(4), 350–371. https://doi.org/10.1016/0022-1031(90)90044-M

Hogarth, R. M., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. *Current Directions in Psychological Science*, *24*(5), 379–385. https://doi.org/10.1177/0963721415591878

Klein, N., & O'Brien, E. (2018). People use less information than they think to make up their minds. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *115*(52), 13222–13227. https://doi.org/10.1073/pnas.1805327115

Le Mens, G., & Denrell, J. (2011). Rational learning and information sampling: On the "naivety" assumption in sampling explanations of judgment biases. *Psychological Review*, *118*(2), 379–392. https://doi.org/10.1037/a0023010

Leslie, L. M. (2019). Diversity initiative effectiveness: A typological theory of unintended consequences. *The Academy of Management Review*, *44*(3), 538–563. https://doi.org/10.5465/amr.2017.0087

Liberman, Z., Woodward, A. L., & Kinzler, K. D. (2017). The origins of social categorization. *Trends in Cognitive Sciences*, *21*(7), 556–568. https://doi.org/10.1016/j.tics.2017.04.004

Lippmann, W. (1922). *Public opinion*. Harcourt, Brace.

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Martinez, J. E., & Todorov, A. (2023). Mapping varied mental representations: The case of representing illegalized immigrants. *Social Cognition*, *41*(6), 507–536. https://joeledmartinez.com/wp-content/uploads/2023/12/martinez-todorov-2023-mapping-varied-mental-representations-the-case-of-representing-illegalized-immigrants.pdf

Meiser, T., & Hewstone, M. (2010). Contingency learning and stereotype formation: Illusory and spurious correlations revisited. *European Review of Social Psychology*, *21*(1), 285–331. https://doi.org/10.1080/10463283.2010.543308

Montepare, J. M., & Zebrowitz, L. A. (1998). Person perception comes of age: The salience and significance of age in social judgments. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 30, pp. 93–161). Academic Press. https://doi.org/10.1016/S0065-2601(08)60383-4

Onyeador, I. N., Hudson, S. K. T., & Lewis, N. A., Jr. (2021). Moving beyond implicit bias training: Policy insights for increasing organizational diversity. *Policy Insights From the Behavioral and Brain Sciences*, *8*(1), 19–26. https://doi.org/10.1177/2372732220983840

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

Oosterhof, N. N., & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion*, *9*(1), 128–133. https://doi.org/10.1037/a0014520

Over, H., & Cook, R. (2018). Where do spontaneous first impressions of faces come from? *Cognition*, *170*, 190–200. https://doi.org/10.1016/j.cognition.2017.10.002

Prager, J., Krueger, J. I., & Fiedler, K. (2018). Towards a deeper understanding of impression formation—New insights gained from a cognitive-ecological perspective. *Journal of Personality and Social Psychology*, *115*(3), 379–397. https://doi.org/10.1037/pspa0000123

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*(4), 741–763. https://dash.harvard.edu/bitstream/handle/1/3207711/Sidanius_SocialDominanceOrientation.pdf

Rhodes, M., & Baron, A. (2019). The development of social categorization. *Annual Review of Developmental Psychology*, *1*(1), 359–386. https://doi.org/10.1146/annurev-devpsych-121318-084824

Rich, A. S., & Gureckis, T. M. (2018). The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology: General*, *147*(11), 1553–1570. https://doi.org/10.1037/xge0000466

Robinson, J., Rosenzweig, C., Moss, A. J., & Litman, L. (2019). Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. *PLOS ONE*, *14*(12), Article e0226394. https://doi.org/10.1371/journal.pone.0226394

Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Individual Differences*, *50*(1), 90–94. https://doi.org/10.1016/j.paid.2010.09.004

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion*, *9*(2), 260–264. https://doi.org/10.1037/a0014681

Shen, X., & Ferguson, M. J. (2021). How resistant are implicit impressions of facial trustworthiness? When new evidence leads to durable updating. *Journal of Experimental Social Psychology*, *97*, Article 104219. https://doi.org/10.1016/j.jesp.2021.104219

Sherman, J. W., Macrae, C. N., & Bodenhausen, G. V. (2000). Attention and stereotyping: Cognitive constraints on the construction of meaningful social impressions. *European Review of Social Psychology*, *11*(1), 145–175. https://doi.org/10.1080/14792772043000022

Skinner-Dorkenoo, A. L., George, M., Wages, J. E., III, Sánchez, S., & Perry, S. P. (2023). A systemic approach to the psychology of racial bias within individuals and society. *Nature Reviews Psychology*, *2*(7), 392–406. https://doi.org/10.1038/s44159-023-00190-z

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science*, *21*(3), 349–354. https://doi.org/10.1177/0956797610362647

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton University Press.

Todorov, A., & Oh, D. (2021). The structure and perceptual basis of social judgments from faces. In B. Gawronski (Ed.), *Advances in experimental social psychology* (Vol. 63, pp. 189–245). Academic Press. https://doi.org/10.1016/bs.aesp.2020.11.004

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

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

Turner, J. C., Brown, R. J., & Tajfel, H. (1979). Social comparison and group interest in ingroup favouritism. *European Journal of Social Ppsychology*, *9*(2), 187–204. https://doi.org/10.1080/01419870500224349

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Uddenberg, S., Thompson, B. D., Vlasceanu, M., Griffiths, T. L., & Todorov, A. (2023). Iterated learning reveals stereotypes of facial trustworthiness that propagate in the absence of evidence. *Cognition*, *237*, Article 105452. https://doi.org/10.1016/j.cognition.2023.105452

Vélez, N., & Gweon, H. (2020). *Preschoolers use minimal statistical information about social groups to infer the preferences and group membership of individuals* [Conference session]. Proceedings of the 42nd Annual Conference of the Cognitive Science Society, Austin, TX, United States.

Verosky, S. C., & Todorov, A. (2010). Generalization of affective learning about faces to perceptually similar faces. *Psychological Science*, *21*(6), 779–785. https://doi.org/10.1177/0956797610371965

Verosky, S. C., & Todorov, A. (2013). When physical similarity matters: Mechanisms underlying affective learning generalization to the evaluation of novel faces. *Journal of Experimental Social Psychology*, *49*(4), 661–669. https://doi.org/10.1016/j.jesp.2013.02.004

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272. https://doi.org/10.1037/0033-295X.114.2.245

Zebrowitz, L. A. (2004). The origin of first impressions. *Journal of Cultural and Evolutionary Psychology*, *2*(1–2), 93–108. https://doi.org/10.1556/JCEP.2.2004.1-2.6

Zebrowitz, L. A. (2017). First impressions from faces. *Current Directions in Psychological Science*, *26*(3), 237–242. https://doi.org/10.1177/0963721416683996

Zebrowitz, L. A., Bronstad, P. M., & Lee, H. K. (2007). The contribution of face familiarity to ingroup favoritism and stereotyping. *Social Cognition*, *25*(2), 306–338. https://doi.org/10.1521/soco.2007.25.2.306

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*(3), 204–223. https://doi.org/10.1207/s15327957pspr0103_2

Zebrowitz, L. A., Fellous, J. M., Mignault, A., & Andreoletti, C. (2003). Trait impressions as overgeneralized responses to adaptively significant facial qualities: Evidence from connectionist modeling. *Personality and Social Psychology Review*, *7*(3), 194–215. https://doi.org/10.1207/S15327957PSPR0703_01

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*(2), 175–189. https://doi.org/10.1037/a0017990

Received November 6, 2023
Revision received August 7, 2024
Accepted August 9, 2024 ■