Accepted for publication at the *Journal of Experimental Social Psychology.* This is a non-final and non-copy edited version of the paper.

Learning to Judge a Book by Its Cover: Rapid Acquisition of Facial Stereotypes

Kao-Wei Chua and Jonathan B. Freeman

New York University

Corresponding Author:
Kao-Wei Chua
Department of Psychology, New York University
Email: kc3644@nyu.edu

Abstract

People are able to quickly and automatically evaluate faces on different traits, such as trustworthiness. There is a growing literature demonstrating that factors such as learning and experience play a role in shaping these judgments. In the current work, we assess the malleability of our trait evaluations by associating arbitrary facial features with trustworthy or untrustworthy behaviors. Across five studies, we demonstrate that this learning can impact trait evaluation and effectively form novel facial stereotypes, which exert effects on evaluations as strong as intrinsic facial trustworthiness. With only a brief training, participants' rapidly acquired novel facial stereotypes, which were activated automatically and early on in processing, and which biased participants' trust behavior and hiring decisions. These results suggest that our trait evaluations of faces are shaped by an implicit learning mechanism that abstracts the co-occurrence between facial features and trait-related behaviors, resulting in the creation of novel facial stereotypes.

Keywords: trait evaluation, implicit learning, face processing, impression formation, stereotyping

Highlights

- A brief training can result in the formation of novel facial stereotypes related to trustworthiness.
- The novel facial stereotypes are highly automatic and activate early on in processing.
- Trait evaluations from faces are malleable and can be formed by learning and experience.

Learning to Judge a Book by Its Cover: Rapid Acquisition of Facial Stereotypes

In our everyday lives, people quickly evaluate others' faces on various trait dimensions, such as trustworthiness or competence (Bar, Neta, & Linz, 2006; Zebrowitz & Montepare, 2008). Such trait evaluations have numerous real-world consequences (for review, see Olivola, Funk, & Todorov, 2014; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015), predicting outcomes such as political success (Todorov, Mandisodza, Goren, & Hall, 2005; Little, Burriss, Jones, & Roberts, 2007), to criminal-sentencing decisions including capital punishment (Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006; Wilson & Rule, 2015). Among the possible traits we perceive from faces, trustworthiness accounts for the bulk of variation and is a proxy for more general evaluation (positive/negative) of faces (Oosterhof & Todorov, 2008; Todorov, Said, Engell, & Oosterhof, 2008). Evaluating trustworthiness from faces is argued to be fundamental and functionally adaptive, consistent with related social cognition models (e.g., evaluating warmth in the stereotype content model; Fiske, Cuddy, Glick, & Xu, 2002; Fiske, Cuddy, & Glick, 2007). Accordingly, tracking others' trustworthiness would help us distinguish friend from foe or those who are likely to help or harm us.

The specific facial features that evoke trustworthy or untrustworthy evaluations and other "facial stereotypes" are now well described (Oosterhof & Todorov, 2008). Facial trustworthiness evaluations emerge as early as 3-4 years of age (Cogsdill, Todorov, Spelke, & Banaji, 2014), and with only 100-ms exposure to faces people make highly consistent evaluations of facial trustworthiness that are strongly correlated across multiple perceivers (Willis & Todorov, 2006). Indeed, trustworthiness is processed so rapidly and automatically that certain brain regions can respond to it even when a face is presented outside conscious awareness (Freeman, Stolier, Ingbretsen, & Hehman, 2014). While research has long emphasized the evolutionarily adaptive

nature of tracking facial trustworthiness, suggesting potential innate mechanisms at play (McArthur & Baron, 1983; Montepare & Dobish, 2003; Oosterhof & Todorov, 2008), recent research has suggested that learning and experience also play a role in shaping trait evaluations from faces (Dotsch, Hassin, & Todorov, 2016; Hehman, Sutherland, Flake, & Slepian, 2017; Sofer et al., 2017; Stolier, Hehman, Keller, Walker, & Freeman, 2018; Stolier, Hehman, & Freeman, 2020).

Here, we are interested in exploring the extent to which trait evaluation from faces can be acquired through implicit learning and derived from experience. It is unlikely that perceivers are learning statistical regularities between facial appearances and targets' own behaviors, as there is little correspondence between facial appearances and one's actual personality or behavior (Rule, Krendl, Ivcevic & Ambady, 2013). However, the manner in which others judge and react to individuals with specific facial appearances tends to be highly consistent across observers (Oosterhof & Todorov, 2008). For instance, if those with furrowed brows are consistently reacted to negatively or if those with upturned mouths are treated in a more positive or trustworthy manner, people may implicitly form associations between those face features and specific traits. Even in the case of young children being able to make trait evaluations that are largely in agreement with adult judgments, it is possible that this ability reflects a statistical learning mechanism that abstracts the co-occurrence between facial features and specific behaviors or traits. For example, if a child consistently observes that people who have a specific pattern of facial features are frequently judged by others as "mean" or that adults do not readily trust such people, this may create an association between those facial features and 'untrustworthy'. With sufficient accumulation of these experiences, such trait associations with

facial appearance would come to reflect preconceived notions derived from the social environment and facial stereotypes could be formed.

Recent research has attempted to use training paradigms that emulate such social learning in order to reduce people's tendency to engage in facial stereotyping. One set of studies demonstrated that a brief counterstereotype learning paradigm (e.g., pairing untrustworthy-appearing faces with trustworthy behaviors) was able to reduce and, in some cases, eliminate the activation and application of facial stereotypes related to trustworthiness (Chua & Freeman, 2020). Whereas this previous work leveraged these learning mechanisms to mitigate the use of existing facial stereotypes, here we aim to test whether a similar training paradigm could result in the rapid creation of completely novel facial stereotypes that affect trustworthiness evaluations. Such findings would highlight the power of learning in driving our facial stereotypes and provide additional insight into how learning, in turn, could reduce or eliminate these stereotypes.

In the current research, we test the hypothesis that implicit learning can result in the creation of novel facial stereotypes. In each study, we associate a novel, arbitrary feature – the width of the sellion (the upper part of the nose bridge) – with trustworthy or untrustworthy behaviors and assess whether this learning has a subsequent impact on face evaluations for new faces that vary in this feature. Previous studies have demonstrated that impression formation based on faces can be informed by covariation in specific features (Hill, Lewicki, Czyzewska, & Schuller, 1990) and preferences for composites faces can be negatively impacted by their resemblance to faces that been paired with aversive stimuli (Jones, Debruine, Little, & Feinberg, 2007). Further, it has long been known that pairing behaviors with faces or providing behavioral information affects trustworthiness evaluations, effectively updating perceivers' impressions of specific exemplars (Todorov & Uleman, 2002, 2003; Verosky, Porter, Martinez, & Todorov,

2018). For instance, learning positive or negative information about a face impacts subsequent evaluations for morphed faces that share a physical similarity, suggesting that affective learning regarding holistic face identities could transfer to other, physically similar faces (Verosky & Todorov, 2010, 2013). Here we used face-behavior pairings to test a distinct kind of updating, whereby perceivers re-associate specific facial appearances themselves with particular traits, unmediated by resemblance to specifically learned face identities. If such statistical learning can occur for specific facial features, this would represent a possible mechanism for the development of facial stereotypes.

If participants are able to rapidly acquire facial stereotypes, the training should result in systematic shifts in face evaluations based on the learned associations with the arbitrary sellion feature. In the present work, we test this rapid acquisition in the context of perceived trustworthiness. We focus on trustworthiness because it is as a proxy for more general face evaluation (positive/negative), and this dimension accounts for more than 60% of the variance in trait judgments of faces (Oosterhof & Todorov, 2008). To test whether the learned associations generalize and are not limited to the specific exemplars they encountered during the learning phase, in each study participants are always tested on a new set of facial identities. Across five studies, we demonstrate that a brief learning phase affects explicit trust-related decisions, both in payment in an economic trust game (Studies 1A/1B) and in hiring decisions of whom to entrust as a financial advisor (Study 2). Further, we show that newly learned associations with sellion width have an early impact on the perceptual process on par with long-established facial trustworthiness using a computer mouse-tracking task (Study 3). We then demonstrate that such a newly acquired facial stereotype even affects automatic evaluations assessed in a more implicit manner via an evaluative priming task (Study 4). Finally, we provide an additional replication to

demonstrate the implicit nature of the learning and rule out alternative explanations (Study 5). Taken together, we provide evidence for a strong role of statistical learning in trait judgments of faces, such that even a very brief training can result in the creation of novel facial stereotypes.

STUDY 1A

In the present study, we test whether arbitrarily learned feature–behavior associations affect the trait evaluation of faces, resulting in novel facial stereotypes. First, participants engage in a short learning phase wherein faces with thin sellions were associated with trustworthy behaviors 80% of the time and faces with wide sellions were associated with untrustworthy behaviors 80% of the time. We then assess the amount of money participants pay in an economic trust game to a new set of faces that vary on sellion width as well as facial trustworthiness. Throughout the article, we refer to the well-characterized set of facial features that drive trustworthiness evaluations as "facial trustworthiness" (Oosterhof & Todorov, 2008), but we do not wish to imply these are the only features related to trustworthiness (indeed, the objective of this research is to demonstrate the malleability of these evaluations from face features). The trust game emulates interaction with human targets and is an explicit trustworthiness measure that has tangible outcomes, as players are told that their payment decisions during the game affects the amount of money that they and other players will receive.

Previous research with trust games have demonstrated that facial trustworthiness predicts payment decisions (Bonnefon, Hopfensitz, & De Neys, 2013), even with brief exposures (~100 ms; De Neys, Hopfensitz, & Bonnefon, 2017) or when the target's past behavioral history is known (Rezlescu, Duchaine, Olivola, & Chater, 2012). Thus, we expect here that trustworthy-looking faces will be paid more money, relative to untrustworthy-looking faces (van 't Wout & Sanfey, 2008; Bonnefon et al., 2013). More importantly, we predict that participants' newly

learned associations with sellion width will also affect payment decisions (i.e., faces with trustworthy-associated sellions will be paid more than untrustworthy-associated sellions). Such results would provide an initial demonstration of the impact of a newly acquired facial stereotypes on face evaluation, even in the presence of otherwise clear trait-related features.

Methods

Participants

One-hundred total participants performed Study 1 on Mechanical Turk for monetary compensation, with 84 in the final sample (mean age = 35.7 years, SD = 11.8 years; 41 male; 60 White, 13 Black, 8 Asian, 3 Other; Hispanic/Latino ethnicity: 5). Without a direct precedent for the current work, we used a target sample size of 88, which was the sample necessary to detect a small-to-medium (d = 0.35) within-subjects effect at 90% power. This standard was used for the samples across all studies. Participants for all studies were paid at a rate of $6 per hour. In these studies, we report all measures, manipulations and exclusions.

A sensitivity analysis (ANOVA, Repeated measures, within factors; $\alpha = 0.05$; one group and two measurements; nonsphericity correction = 1) was conducted using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007). This analysis indicated that with a final sample of n = 84, the minimum detectable effect size was $\eta_p^2 = 0.087$ at 80% power.

Procedure

Using sellion width (the upper part of the nose bridge), we generated wide and narrow variants for each of 40 unique face identities (see Figure 1). Thus, the wide vs. narrow sellion variants were equated on all other perceptual information except sellion width. The 40 unique identities were realistic male faces generated by FaceGen Modeler, which uses the statistical face model described by Blanz & Vetter (1999). Specifically, the target faces were at their population

average for all feature dimensions, in addition to being neutral in emotional expression and at the population mean for trustworthiness (Oosterhof & Todorov, 2008). The sellion width manipulation was conducted in FaceGen Modeler using the sellion slider (-2 *SD* for narrow sellions, +2 *SD* for wide sellions). Half of the 40 identities were presented during the learning phase, and the other half of identities were used in the evaluation phase. Face stimuli in the learning and evaluation phases in this study and the following studies were 280 pixels in height x 280 pixels in width and presented in the center of the browser window. Data were collected online via MTurk, so other display conditions such as viewing distance were uncontrolled.

To ensure the sellion manipulation did not have its own relationship with facial trustworthiness perception (without learning), we recruited 40 independent raters from Mechanical Turk (M = 35.1 years, SD = 10.2 years; 22 male; race: 29 White, 5 Black, 3 Asian, 3 Other; Hispanic/Latino ethnicity: 6), who assessed each of the face stimuli (narrow- and wide-sellion variants of the 40 identities) on trustworthiness using a 7-point Likert scale. There was high inter-rater agreement (intraclass correlation coefficient [ICC] = 0.94). Using mean ratings of trustworthiness for each target, an items-based paired t-test (comparing mean ratings for the 40 identities) revealed no significant difference in perceived trustworthiness between narrow- vs. wide-sellion targets, $t(39) = 1.29$, $p = 0.21$, $d = 0.29$; a raters-based paired t-test (comparing mean ratings of the 40 raters) also revealed no significant difference, $t(39) = 1.25$, $p = 0.22$, $d = 0.27$. These results confirm that the feature manipulation did not have an impact on perceived trustworthiness. Moreover, additional analyses confirmed the feature manipulation did not have an effect on perceived competence or dominance as well (see Supplementary Materials).

In Study 1, the first task consisted of a learning phase that was framed to participants as a test of memory for individual faces. Participants viewed 20 target faces paired with brief one-

sentence behavioral descriptions, and they were instructed to memorize the face–behavior pairings, as they would be important in a later portion of the study. Half of the 20 target faces had narrow sellions and the remaining half had wide sellions. Participants were never presented with narrow and wide sellion variants of the same face identity. The behaviors were taken from previous research, and testing showed that the trustworthy and untrustworthy behaviors were perceived accordingly and matched on intensity (Chua & Freeman, 2020).

The faces with narrow sellions were paired with trustworthy behaviors (e.g., "Volunteered at a homeless shelter") 80% of the time and the faces with wide sellions were paired with untrustworthy behaviors (e.g., "Threw a rock at a neighbor's window") 80% of the time (see the Supplementary Table 1 for full list of behaviors). Participants viewed each slide with face–behavior information at their own pace, and trials were presented in a randomized order. There was an error warning that remained on screen for 2000 ms if participants clicked through a slide in under 500 ms, so as to encourage participants to attend to the face–behavior pairings. In total, each of the target face–behavior pairings was shown three times, resulting in 60 total trials for the learning phase.

Figure 1. Examples of faces varying in sellion width. The faces in the top row have sellions that are two standard deviations wider than the population mean, and the faces on the bottom row have faces that are two standard deviations narrower than the population mean.

Following the learning task, participants completed an ostensibly unrelated task. This task was an economic trust game that involved decisions to allocate money to other human players, as used in several previous studies (Berg, Dickhaut, & McCabe, 1995; van 't Wout & Sanfey, 2008; Chang, Doll, van 't Wout, Frank, & Sanfey, 2010). A number of steps were taken to improve the realism of the task. First, participants were told that they and the other players were playing for real money and the instructions urged them to make decisions to maximize profits. Second, before the game began, participants selected an avatar from a large set of faces that would represent them to other players, and they were told that the other human players would also be represented by similar avatars. Third, the time in between each round of the trust game began with a delay of random length, ranging from 1000 to 10000 ms, to simulate a delay wherein the computer was finding a match for the participant. In actuality, there were no other human players and every participant was presented with a fixed set of faces presented in a randomized order.

At the beginning of the trust game, participants were told that they were randomly assigned to be Player 1, who was designated as the “giver.” For each trial, they would receive $1.00 and had the opportunity to give some of that money to another player (five options: $0.00, $0.25, $0.50, $0.75, $1.00). Whatever amount given would then be tripled, and their partner in the game would have the opportunity to return as much or as little money as they wanted to the participant. For example, if the participant gave the full $1.00, the other player would then have $3.00 and could give whatever portion of that amount back to the participant. However, if the

participant found the other player's avatar to be untrustworthy, for example, they could give less money and keep the remaining amount for themselves for that particular round.

The trust game involved 32 total trials or rounds. On 16 of the rounds, participants were paired with 8 unique male partners twice. The 8 unique male face identities were drawn from the 20 identities designated for the evaluation phase, as described earlier. These 8 partners comprised a 2 (sellion width: wide vs. narrow) x 2 (facial trustworthiness: untrustworthy vs. trustworthy) design. The remaining 16 rounds consisted of 8 female avatars presented twice with no systematic variability in sellion width; these faces were included as distractors to reduce the transparency of the task. These filler female avatars were of no interest and were created using the same methods described earlier. Critically, the set of faces in the trust game consisted of distinct identities from the training phase, so that we could test whether the facial feature–behavior associations would transfer to novel targets that shared those same, diagnostic face features.

Five attention checks were interspersed between trust game rounds to ensure participants maintained attention, wherein participants were instructed to press a number (1-5). After the trust game, participants learned that there would be no test of face memory as purported in the learning phase. Instead, they provided demographic information and were debriefed about the study aims.

Results and Discussion

Before conducting analyses, 9 participants were removed for failing attention checks interspersed throughout the study. Another 6 participants were removed for triggering error warnings for excessive speed (< 500 ms) on the majority of trials during the learning phase (indicating that they were not closely reading the behaviors associated with each face). Data and

analysis code for all studies are available at https://osf.io/sba6j/?view_only=0b5d5833463141e2b42ce3cd0b969396.

Payment in the trust game was submitted to a 2 (Sellion Width: Trustworthy vs Untrustworthy) x 2 (Facial Trustworthiness: Trustworthy vs Untrustworthy) repeated-measures ANOVA. Expectedly, there was a significant effect of facial trustworthiness, $F(1,83) = 42.00$, $p < 0.0001$, $\eta_p^2 = 0.34$, with more payment allocated to faces that had higher facial trustworthiness. More importantly, there was also a significant effect of sellion width, $F(1,83) = 75.70$, $p < 0.0001$, $\eta_p^2 = 0.48$, such that more money was given to faces with trustworthy vs. untrustworthy sellions. Interestingly, there was a significant interaction between sellion width and facial trustworthiness, $F(1,83) = 5.85$, $p = 0.018$, $\eta_p^2 = 0.07$ (see Figure 2). This interaction arose due to an additive effect, whereby having both an untrustworthy sellion and an untrustworthy face led participants to allocate an especially low amount of payment. While faces with trustworthy sellions were given less money when the face was untrustworthy as opposed to trustworthy, $F(1,83) = 17.15$, $p < 0.0001$, $\eta_p^2 = 0.17$, this difference due to facial trustworthiness was considerably more pronounced for faces with untrustworthy sellions, $F(1,83) = 42.99$, $p < 0.0001$, $\eta_p^2 = 0.34$.

To alleviate the concern that these results could have been spuriously produced by shifts in other trait dimensions such as competence or dominance, we statistically controlled for the faces' competence and dominance in a trial-by-trial manner using a multi-level regression framework (see Supplementary Materials for details). None of the reported results were substantially changed after additionally controlling for facial competence and facial dominance, including their interactions with sellion width and facial trustworthiness (Supplementary Tables 2 and 3).

Figure 2. Violin plots representing the probability densities for payment by facial trustworthiness and sellion width for Study 1A. Error bars represent the standard error of the mean.

In Study 1A, we established a relationship between learned feature–trait pairings and subsequent behavior in an economic trust game, with a novel facial stereotype exerting a comparable effect as the trustworthiness of the face. Both the novel sellion stereotype and facial trustworthiness exerted significant effects on payment. Additionally, we demonstrate that these factors had an additive effect: having both an untrustworthy-looking face and an untrustworthy-associated sellion elicited a pronounced decrease in payment. However, one concern of the present study is that we trained on only one feature–behavior association (narrow = trustworthy, wide = untrustworthy). To address any potential confounds in the feature-behavior assignment, we seek to replicate these effects in Study 1b with the reversed associations.

STUDY 1B

It is possible that the effects in payment in Study 1A are somehow limited to the feature–behavior associations used for that study, namely that narrow sellions were learned to be trustworthy and wide sellions learned to be untrustworthy. In Study 1B, we aim to replicate the effects with the counterbalanced associations (narrow sellion = untrustworthy, wide sellion = trustworthy).

Methods

Participants

In Study 1B, we targeted an equivalent sample size as in Study 1A. One hundred participants performed Study 1B on Mechanical Turk, with a final sample of 83 participants (mean age = 33.4 years, SD = 11.5 years; 41 Male, 57 White, 14 Black, 8 Asian, 4 Other; Hispanic/Latino ethnicity: 7). Participants received monetary compensation.

A sensitivity analysis (ANOVA, Repeated measures, within factors; $\alpha = 0.05$; one group and two measurements; nonsphericity correction = 1) was conducted, and this analysis indicated that with a sample size of n = 83 at 80% power, the minimum detectable effect size was $\eta_p^2 = 0.089$.

Procedure

The procedure was identical to that of Study 1A, except that for the learning phase, the association between sellion width and valenced behaviors was reversed. That is, faces with wide sellions were associated with trustworthy behaviors and faces with narrow sellions were associated with untrustworthy behaviors. The face stimuli used during the learning phase and the trust game were the same as in Study 1A.

Results and Discussion

Before conducting analyses, 12 participants were removed for failing attention checks interspersed throughout the study. Another 5 participants were removed for triggering the minimum timeouts on the majority of trials during the learning phase. The final analyses involved 83 participants.

Payment in the trust game was submitted to a 2 (Sellion Width: Trustworthy vs Untrustworthy) x 2 (Facial Trustworthiness: Trustworthy vs. Untrustworthy) repeated measures ANOVA. There was the predicted main effect of facial trustworthiness, $F(1,82) = 8.39$, $p = 0.005$, $\eta_p^2 = 0.09$, with more money paid to trustworthy faces. More critically, there was a main effect of sellion width, $F(1,82) = 17.01$, $p < 0.0001$, $\eta_p^2 = 0.17$, with more money paid to faces with trustworthy as opposed to untrustworthy sellions. The interaction between sellion width and facial trustworthiness was not significant, $F(1,82) = 0.88$, $p = 0.35$, $\eta_p^2 = 0.01$. See Figure 3. As in Study 1A, these effects still held when including the influence of facial competence and facial dominance (and their interactions with sellion width and facial trustworthiness) as covariates (Supplementary Tables 4 and 5).

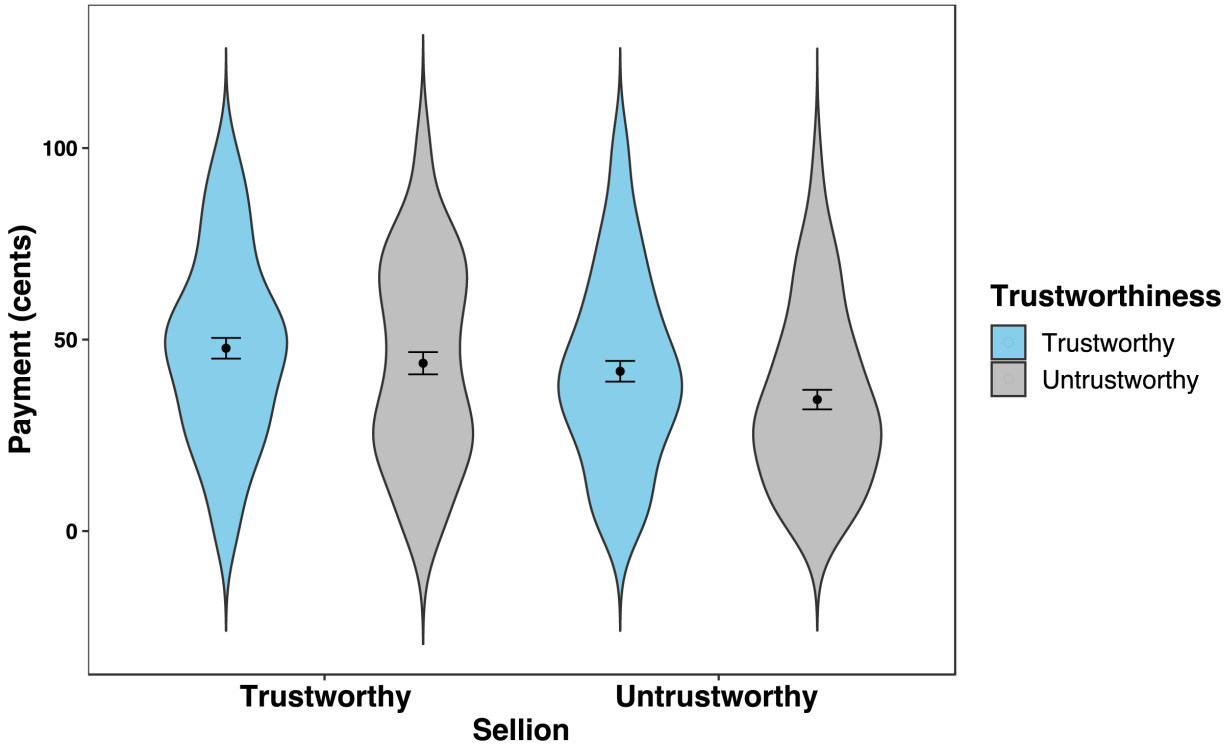

Figure 3. Violin plot representing the probability densities for payment by trustworthiness and sellion-associated behavior for Study 1B. Error bars represent the standard error of the mean.

Here, we replicated the main effects in Study 1A wherein a novel stereotype involving sellion width affected payment in an economic trust game. Both sellion width and the "intrinsic" trustworthiness of faces exerted an effect on payment decisions. With Study 1B, we demonstrate that the effects observed in Study 1A were not an artifact of the specific assignment of sellion width and trustworthy behaviors by demonstrating that these effects are observed with either feature–behavior configuration. In all subsequent studies, which sellion is associated with which valenced behavior is counterbalanced across participants.

STUDY 2

In Study 1, we showed that an arbitrary, newly learned facial stereotype can impact to what extent participants trusted targets in the context of an economic game. In Study 2, we extended these results to another consequential domain in which facial stereotyping can readily

bias outcomes, namely hiring and candidate selection processes. We implement the same training regimen as in Study 1, but we follow it with a line-up choice regarding which target individual participants would like to select as their financial advisor and to entrust to handle their finances. We hypothesize that participants' selection of their financial advisor will be shaped by both the newly learned facial stereotype related to sellion width as well as facial trustworthiness.

Methods

Participants

One-hundred nineteen total participants were recruited from Mechanical Turk in exchange for monetary compensation, with a final sample of one-hundred twelve (mean age = 32.9 years, SD = 11.3 years, 54 male, 89 White, 10 Black, 9 Asian, 4 Other; Hispanic/Latino ethnicity: 15).

A sensitivity analysis (chi-squared goodness of fit test: contingency tables; $\alpha = 0.05$, 1 degree of freedom) was conducted. With a sample of 112, this analysis yielded a minimum detectable effect size of $w = 0.26$ at 80% power.

Procedure

As in Study 1, the study took place in two phases, with a learning phase and the lineup choice. The learning phase was identical to Studies 1A/1B, except the association between sellion width and valenced behavior was counterbalanced across participants. Following the learning phase, participants were instructed that the second part of the study would examine how people make hiring decisions. Participants read a short paragraph about how the recent financial crisis was brought on by financial advisors who made risky investment decisions and that the purpose of the study was to examine how people made hiring decisions based on appearances,

consistent with previous studies examining the role of facial appearance in biasing hiring decisions (e.g., Hehman, Flake, & Freeman, 2015).

Following the instructions, participants were shown an array of the faces of four target individuals and instructed to indicate which individual they would choose to handle their personal finances. Each of the four target individuals' faces represented each condition of interest: 1) trustworthy face, trustworthy sellion; 2) trustworthy face, untrustworthy sellion; 3) untrustworthy face, trustworthy sellion; 4) untrustworthy face, untrustworthy sellion. These four unique identities were drawn from the 20 identities designated for the evaluation phase, which were independent from those used in the learning phase (see Methods of Study 1A). There were 16 potential arrays, one of which was chosen at random for each participant. Participants were then debriefed and demographic information was collected.

Results and Discussion

Seven participants were removed because a majority (>50%) of trials in the learning phase incurred error warnings for excessive speed (<500 ms). The final sample for the analysis included 113 participants.

A chi-squared goodness-of-fit test was performed to determine whether the choice of financial advisor was equally preferred (chance performance = 25%). The choice of financial advisor was not equally distributed across the four faces, χ^2 (4,112) = 62.5, $p < 0.0001$, $w = 5.91$. Participants most often selected targets with a trustworthy sellion and trustworthy face, next followed by a trustworthy sellion and untrustworthy face, an untrustworthy sellion and trustworthy face, and least often selected those with an untrustworthy sellion and untrustworthy face. Two-sample proportion tests showed that participants were significantly more likely to select targets with trustworthy faces if they also had a trustworthy sellion (52.6%, 95% CI:

43.2% - 62.1%) as opposed to untrustworthy sellion (10.7%, 95% CI = 3.3% - 18.3%), $\chi^2 = 43.6$, $p < 0.0001$, $w = 4.12$; they were also more likely to select targets with untrustworthy faces if they also had a trustworthy sellion (31.3%, 95% CI: 21.7% - 40.8%) as opposed to untrustworthy sellion (5.3%, -1.1% - 11.8%), $\chi^2 = 23.4$, $p < 0.0001$, $w = 2.21$.

Thus, regardless of a trustworthy or untrustworthy facial appearance, the newly learned facial stereotype related to sellion width strongly impacted participants' decisions regarding who they wished to hire as their financial advisor. These results extend those of Studies 1A/1B, demonstrating rapidly acquired facial stereotypes in the context of both payment allocations in an economic trust game as well as hiring decisions.

STUDY 3

In Studies 1A/1B and 2, we demonstrate that a brief learning period resulted in a novel facial stereotype that affected evaluations based on payment in an economic trust game and in choosing a personal financial advisor. However, while these newly learned associations clearly biased the outcomes of such explicit evaluations, it is unclear to what extent they activated automatically or biased evaluations early on in processing. Facial trustworthiness has long been known to activate automatic evaluations and is processed very rapidly (Willis & Todorov, 2006; Engell, Haxby, & Todorov, 2007; Freeman et al., 2014). Can even a new facial stereotype acquired via minimal training have such pronounced effects on the evaluative process as "intrinsic" facial trustworthiness?

The results of Studies 1A/1B and 2 show that the newly learned associations related to sellion width have a strong effect on evaluative decisions; indeed, the effect is on par with long-established associations related to facial trustworthiness. Here, we tested how early in processing a newly learned facial stereotype can exert its bias. Facial trustworthiness has long been known

to have early and automatic effects on evaluation. In the current study, we tested whether newly formed facial stereotype associations bias early processing on par with facial trustworthiness, or only exert effects later in processing given that they are only recently formed.

Current models of face-based social perception argue that multiple facial features simultaneously weigh in on partially-active representations (e.g., trustworthy), which dynamically integrate evidence over time until stabilizing on a given response, such as whether a face is "trustworthy" vs. "untrustworthy" (Freeman & Ambady, 2011; Freeman & Johnson, 2016). These predict that, in cases of conflicts (e.g., untrustworthy face with trustworthy sellion), both features will drive the evaluative process in parallel (e.g., "untrustworthy" and "trustworthy" simultaneously activated) continuously over time until a given evaluation wins out. This would only occur, however, if both features are strongly associated with a response and activated sufficiently early in processing. By contrast, if one feature is processed earlier in time (e.g., untrustworthy face) and another feature processed later (e.g., trustworthy sellion), then a stage-based sequence would be expected (e.g., "untrustworthy" activates first, which is then replaced by "trustworthy"). Because facial trustworthiness has strong associations present via long-term learning as early as 4 years of age (Cogsdill et al., 2014), it is possible that this feature may drive processing early on, with newly learned associations (e.g., sellion width) only later able to exert their impact. These kinds of stage-based dynamics have been observed previously, e.g., gender categorization of long-haired men or short-haired women (Martin & Macrae, 2007; Freeman, 2014). Here, we predict that, even though sellion width's evaluative associations are learned only recently during a brief training, it will have an early impact on processing on par with long-established facial trustworthiness.

To examine this, we use a mouse-tracking task to examine how both facial trustworthiness and sellion width drive the time-course of trustworthiness evaluations. During mouse-tracking, participants begin a trial by pressing a start button at the bottom-center of the screen, after which a face is presented. They must then rapidly select a response in either top corner of the screen (e.g., "trustworthy" vs. "untrustworthy"), during which their mouse trajectory is recorded. When the sellion width is incongruent with facial trustworthiness (e.g., trustworthy face with untrustworthy sellion), we predict participants' hand movements to exhibit a partial, simultaneous attraction toward the opposite response (e.g., "untrustworthy") continuously over time due to the early and parallel impact of the sellion width association. In other words, at every moment during the processing stream, participants' response trajectory would always be reflecting some weighted combination of both facial trustworthiness and the sellion width association simultaneously.

An alternative possibility is that when sellion width is incongruent with facial trustworthiness, participants initially head directly toward the response associated with facial trustworthiness (e.g., "trustworthy") and then only once the sellion width association becomes available in the processing stream, they redirect their trajectory straight to the opposite response associated with sellion width (e.g., "untrustworthy"). If this alternative trajectory pattern were observed, it would suggest that long-established facial trustworthiness is processed first, with the more newly learned sellion width association only coming online after facial trustworthiness.

The mouse-tracking technique has long been used to dissociate these different patterns of temporal dynamics (for review, Freeman, 2018), and these two alternative possibilities make different empirical predictions. In testing the effects of sellion width and facial trustworthiness on mouse trajectories' spatial attraction toward the opposite response, both possibilities predict a

significant interaction effect: When sellion width and facial trustworthiness are incongruent (i.e., trustworthy faces with untrustworthy sellions and untrustworthy faces with trustworthy sellions), regardless of participants' ultimately selected response, there should be greater attraction to the opposite response due to both cues being processed. However, the former possibility (sellion width is processed as early as trustworthiness) predicts a modest amount of attraction to the opposite response that is unimodally distributed across incongruent trials. The latter alternative (sellion width is processed after trustworthiness) predicts either 1) an extremely high amount of attraction to the opposite response that is unimodally distributed across incongruent trials; or 2) a modest amount of attraction to the opposite response that is bimodally distributed across incongruent trials (Freeman, 2018; Freeman & Dale, 2013). This is because, if on all incongruent trials, participants head straight toward an initial response (due to facial trustworthiness) then subsequently correct themselves, redirecting toward a second response (due to sellion width), this will produce an extremely high amount of attraction that is present on all incongruent trials and thus unimodally distributed. A more realistic variant of this account would predict that only for some incongruent trials participants engage in a stage-based correction process, while on the remainder of incongruent trials they simply head straight toward one response (but never heading to both responses at the same time); if true, this would average out to a modest amount of average attraction across incongruent trials but one that is bimodally distributed (due to one subpopulation of trials having low attraction and another subpopulation having high attraction) (Freeman, 2018; Freeman & Dale, 2013).

We disentangle these two possibilities by examining participants' mouse-trajectory attraction effect for incongruent (trustworthy faces with untrustworthy sellions and untrustworthy faces with trustworthy sellions) vs. congruent (trustworthy faces with trustworthy sellions and

untrustworthy faces with untrustworthy sellions) targets in tandem with this effect's unimodal vs. bimodal distribution. As we predict the newly learned sellion width association to be processed as early as long-established facial trustworthiness, we predict incongruent targets to elicit a modest amount of spatial attraction to the opposite response that is unimodally distributed.

Methods

Participants

One-hundred participants were recruited from Mechanical Turk in exchange for monetary compensation, with a final sample of 90 participants (mean age = 34.9 years, SD = 9.7 years; 49 male, 69 White, 8 Black, 6 Asian, 7 Other; Hispanic/Latino ethnicity: 10). Participants were eligible to participate only if they were using a mouse and not a trackpad.

A sensitivity analysis (ANOVA, Repeated measures, within factors; $\alpha = 0.05$; one group and two measurements; nonsphericity correction = 1) was conducted using G*Power (Faul et al., 2007). This analysis indicated that with a sample of n = 90 at 80% power, the minimum detectable effect size was $\eta_p^2 = 0.083$.

Procedure

Participants underwent a learning phase identical to Studies 1A/1B and 2, where they learned the association between narrow vs. wide sellions and trustworthy vs. untrustworthy behaviors (counterbalanced across participants).

Following the learning phase, participants were instructed about a second part of the study that involved categorizing different faces as trustworthy or untrustworthy. On each trial, a start button appeared at the bottom center of the screen, which was replaced by a face when clicked. Faces were presented in a randomized order, and "trustworthy" and "untrustworthy" response options appeared at either top corner of the screen. The placement of the categories on

the left vs. right was counterbalanced for each participant. Mouse-trajectory data were recorded during the categorization process using a Javascript-based implementation of MouseTracker software (Freeman & Ambady, 2010).

Following practice trials, participants proceeded to the main mouse-tracking task. Face stimuli were 8 unique identities presented five times each, resulting in 40 trials total for each participant. The faces could have wide or narrow sellions and trustworthy or untrustworthy faces (+2 and -2 SD), resulting in 4 total conditions and two unique identities within each condition. These unique identities were drawn from the 20 identities designated for the evaluation phase, which were independent from those used in the learning phase (see Methods of Study 1A). Upon completion of the mouse-tracking task, participants were debriefed and demographic information was collected. A set of five attention check trials were also included, where participants were explicitly told which response to select; this helped ensure participants were paying attention during the duration of the task.

Results and Discussion

As in previous studies, six participants were removed for having a majority of trials in the learning phase proceed with excessive speed. Another four participants were removed for always choosing the same option (trustworthy or untrustworthy) during the mouse-tracking section and/or inputting the incorrect response during attention check trials. This left 90 participants for final analysis.

Mouse-trajectory preprocessing followed standard procedures (Freeman & Ambady, 2010). To permit comparison across trials, trajectories for all trials were remapped such that the selected response was on the right and the unselected response was on the left. Trajectories were rescaled into a standard x, y coordinate space: top left at [−1, 1.5] and bottom right at [1, 0],

leaving the start position of the mouse at [0, 0], and were normalized (linearly interpolated) into 101 time steps (100 time bins) to permit averaging of their full length across multiple trials. To index the degree of spatial attraction toward the opposite response, for each trial, the area under the curve (AUC) was computed: the geometric area between the observed trajectory and an idealized response trajectory (a straight line between the trajectory's start and endpoints).

AUC was submitted to a 2 (Sellion Width: Trustworthy vs Untrustworthy) x 2 (Facial Trustworthiness: Trustworthy vs. Untrustworthy) repeated-measures ANOVA. There were no main effects of sellion width, $F(1,89) = 1.08$, $p = 0.30$, $\eta_p^2 = 0.01$, or facial trustworthiness, $F(1,89) = 1.50$, $p = 0.23$, $\eta_p^2 = 0.02$. More critically, consistent with our predictions, there was a significant interaction (Figure 4), $F(1,89) = 22.70$, $p < 0.0001$, $\eta_p^2 = 0.20$. For trustworthy sellions, there was greater attraction toward the opposite response when facial trustworthiness was low (incongruent) compared to when facial trustworthiness was high (congruent), $F(1,89) = 5.49$, $p = 0.02$, $\eta_p^2 = 0.06$. When the sellion was untrustworthy, however, there was greater attraction when facial trustworthiness was high (incongruent) compared to when facial trustworthiness was low (congruent), $F(1,89) = 12.59$, $p = 0.001$, $\eta_p^2 = 0.12$. The critical interaction between sellion and trustworthiness held even when statistically controlling for facial competence and facial dominance (and their interactions with sellion width and facial trustworthiness) (Supplementary Tables 6 and 7).

Figure 4. Violin plots representing probability density for the area under the curve (AUC) for trustworthy categorizations in Study 3. Error bars represent the standard error of the mean.

As mentioned, a higher AUC for incongruent relative to congruent conditions is consistent with both the possibility that sellion width is processed as early as facial trustworthiness or the possibility that it is processed only after facial trustworthiness. A seen in Figure 4, the differences in AUC between incongruent vs. congruent conditions were strong but modest, rather than extreme, and on casual inspection appear to be unimodally distributed. To formally distinguish between a unimodal vs. bimodal distribution, previous mouse-tracking studies and methodological simulations have shown that inspecting incongruent conditions' AUC distribution for bimodality (i.e., two modes of high AUC and low AUC) can reliably detect such a stage-based pattern of results (i.e., trajectories head first toward the trustworthiness-associated response, then afterward head toward the sellion-associated response), and that the Hartigan's Dip Statistic is the optimal measure of bimodality in this context (Hartigan &

Hartigan, 1985; Freeman & Dale, 2013). Examining the incongruent conditions, neither the trustworthy sellion – untrustworthy face ($D = 0.009$, $p = 0.91$) nor the untrustworthy sellion – trustworthy face conditions (0.008, $p = 0.96$) significantly departed from a unimodal distribution, ruling out the possibility of a stage-based process and suggesting that sellion width had an early and parallel impact with facial trustworthiness.

This pattern of mouse trajectories indicates that when the newly learned sellion cue was incongruent with facial trustworthiness, participants' hand movements elicited a partial and parallel attraction toward both the "trustworthy" and "untrustworthy" response options. Such a result suggests that both facial trustworthiness and sellion width were driving the real-time evaluation process simultaneously over time, and that sellion width had an early impact on par with long-established facial trustworthiness cues. These results demonstrate that even a short learning period can result in a novel facial stereotype that exerts an early impact on evaluation processing, weighing in on real-time evaluations in parallel with more entrenched facial trustworthiness features.

STUDY 4

The previous studies show that a brief training resulted in a novel facial stereotype that can impact explicit judgments of trustworthiness and has an early impact on evaluative processing in parallel with "intrinsic" facial trustworthiness features. One open question is to what extent this novel facial stereotype is automatized and has an impact on more implicit measures of evaluation. This would cast doubt on the possibility that the previous effects are mere artifacts of demand characteristics, highlight the power of brief feature-behavior learning, and demonstrate the malleability of facial trustworthiness evaluation.

To measure the automaticity of the evaluation of the newly learned sellion cues, here we use an evaluative priming task (Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Giner-Sorolla, Garcia, & Bargh, 1999). In evaluative priming tasks, a prime, usually a picture or word, is briefly presented and is followed by an evaluation of a target word (e.g., joy, poison) as positive or negative. Primes that are implicitly evaluated as positive should facilitate the response times (RTs) to positive words while primes implicitly evaluated as negative should facilitate RTs to negative words (Fazio, 2001).

Methods

Participants

One-hundred participants completed Study 4 on Mechanical Turk for monetary compensation. The final sample consisted of 84 participants (mean age = 35.4 years, SD = 8.4 years; 40 male, 59 White, 11 Black, 7 Asian, 7 Other; Hispanic/Latino ethnicity: 10). A sensitivity analysis (ANOVA, Repeated measures, within factors; $\alpha = 0.05$; one groups and two measurements; nonsphericity correction = 1) was conducted, and this analysis indicated that with a sample size of n = 84 at 80% power, the minimum detectable effect size was $\eta_p^2 = 0.088$.

Procedure

Following the learning phase as in previous studies (with counterbalanced sellion associations), participants were told that they would be performing a word recognition task that would be a test of their language abilities. On each trial, participants were shown a fixation cross for 500 ms, followed by a prime face shown for 200 ms, followed by a 100 ms blank screen, followed by target word that appeared in the center of the screen until a response. The procedure follows previous studies achieving robust facilitation effects by the prime (Fazio, Sanbonmatsu,

Powell, & Kardes, 1986; Hermans, Houwer, & Eelen, 1994). The task was to classify the target word as positive or negative as quickly and accurately as possible by key press ("S" for positive", "K" for negative). The face stimuli were identical to those in the previous studies, consisting of faces with wide or narrow sellions and trustworthy and untrustworthy faces, with five unique identities for each condition, resulting in 20 total prime faces. These unique identities were drawn from the 20 identities designated for the evaluation phase, which were independent from those used in the learning phase (see Methods of Study 1A). The target words were adjectives that had positive (e.g., "Good", "Kind", "Pleasant") and negative ("Bad", "Mean", "Unfriendly") valences. There were 10 target words each for positive and negative valences. Each prime face (20 faces) was paired with each target word once, resulting in 400 trials total.

Results and Discussion

As in the previous studies, six participants were removed for having the majority of learning phase trials proceed with excessive speed. Another ten participants were removed for being at or below chance accuracy for the word evaluation decision (50%). This left 84 participants for the final analysis.

For the evaluative priming data, we removed incorrect responses (3% of trials removed) as well as trials with reaction times faster than 250 ms and slower than 3000 ms (4% of trials removed). RT difference scores [negative words – positive words] were calculated. In this case, a positive difference score reflects greater facilitation for categorizing positive words (i.e., a positive evaluation) and a negative difference score reflects greater facilitation for categorizing negative words (i.e., a negative evaluation). The RT difference scores were submitted to a 2 (Sellion Width) x 2 (Facial Trustworthiness) repeated-measures ANOVA. As expected, there was a significant main effect of facial trustworthiness on reaction time facilitation, $F(1,83) =$

11.49, $p = 0.001$, $\eta_p^2 = 0.12$. Critically, there was also a main effect of sellion width, $F(1,83) = 16.74$, $p < 0.0001$, $\eta_p^2 = 0.17$. Consistent with the pattern of results found in Study 1, the effect size of sellion width ($\eta_p^2 = 0.17$) was greater than that of facial trustworthiness ($\eta_p^2 = 0.12$). There was no interaction between sellion width and facial trustworthiness, $F(1,83) = 0.15$, $p = 0.70$, $\eta_p^2 = 0.002$. These effects held even when statistically controlling for facial competence and facial dominance (and their interactions with sellion width and facial trustworthiness) (Supplementary Tables 8 and 9).

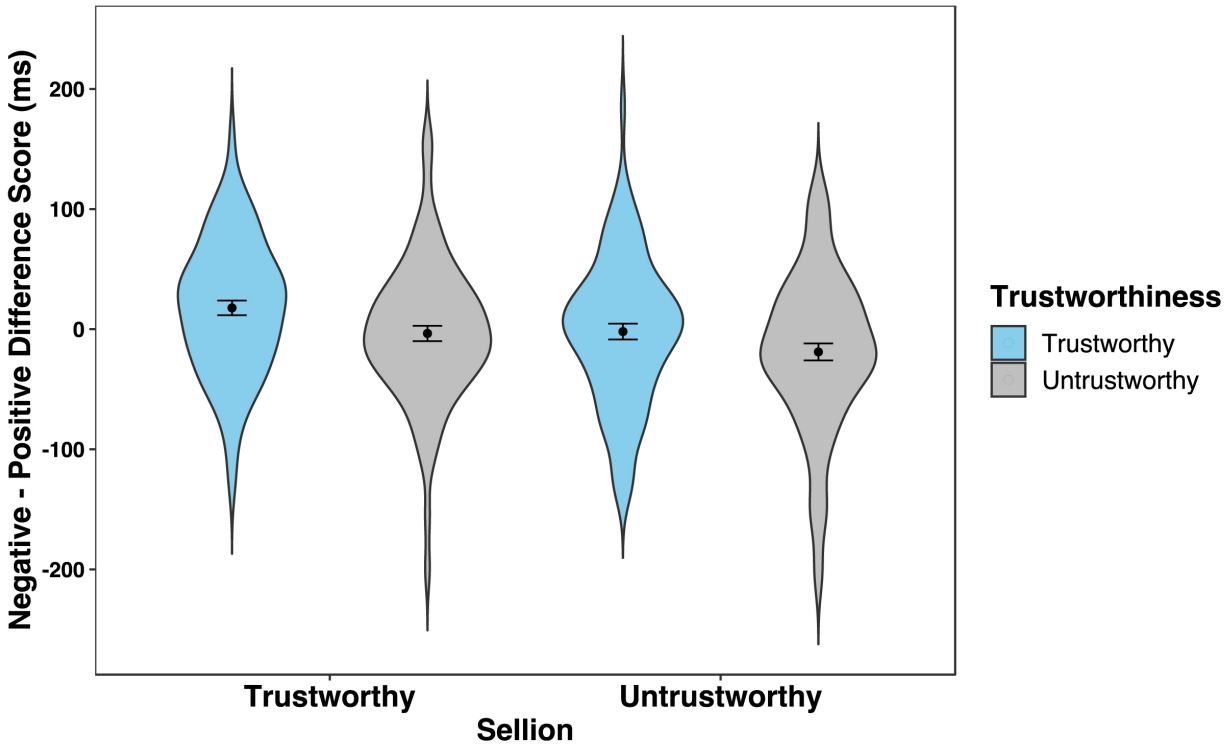

Figure 5. Violin plots representing probability densities for RT facilitation effects for sellion width and facial trustworthiness. RT facilitation is indexed as a difference score for positive – negative words. Facilitation towards positive is registered as a negative value (less than zero) and facilitation towards negative is registered as a positive value (greater than zero). Error bars represent the standard error of the mean.

These results show that novel facial stereotypes derived from a brief learning period were able to not only affect explicit evaluations of faces but also more automatic, implicit evaluations

as well.

STUDY 5

One distinct possibility throughout the previous studies is that participants were explicitly aware of the sellion-width association with trustworthiness and that the shifts in evaluations observed in each study were driven by heuristics or rules guided by this awareness or by demand characteristics. Another potential issue to resolve is the extent to which the effects of the training phase are truly generalizing to the novel exemplars with the learned feature during the evaluation phase. In theory, the faces learned during the training phase could have been conflated with those presented during the evaluation phase (e.g., due to them both being computer-generated and appearing generally similar), which if true, would weaken our claim about how the learning is generalizing and being applied as a novel stereotype.

To address these issues, in Study 5 we conducted a replication of Studies 1A/1B using the economic trust game but added two additional elements. To probe participants' explicit awareness of the sellion-width association, following the task, we asked participants whether they were aware of the sellion-width feature and its mapping to trustworthy/untrustworthy behaviors while making their evaluations. Moreover, to test whether the faces during the learning phase and evaluation phase could have been potentially conflated (thereby casting doubt on any true generalization of the learning), participants performed a surprise face memory task where they were asked to indicate whether faces had previously been presented during the learning vs. the evaluation phases. Thus, Study 5 aims to provide an additional replication of Studies 1A/1B while also ruling out two possible concerns regarding the nature of the learning effects.

Methods

Participants

One-hundred participants were recruited from Mechanical Turk in exchange for monetary compensation, with a final sample of 88 participants (mean age = 34.9 years, SD = 9.7 years; 49 male, 69 White, 8 Black, 6 Asian, 5 Other; Hispanic/Latino ethnicity: 10). Participants were eligible to participate only if they were using a mouse and not a trackpad.

A sensitivity analysis (ANOVA, Repeated measures, within factors; $\alpha = 0.05$; one group and two measurements; nonsphericity correction = 1) was conducted using G*Power (Faul et al., 2007). This analysis indicated that with a final sample of n = 88, the minimum detectable effect size was $\eta_p^2 = 0.084$ at 80% power.

Procedure

The learning phase and the economic trust game were identical to Studies 1A/1B. However, unlike Study 1A (narrow sellion = trustworthy, wide sellion = untrustworthy) and Study 1B (narrow sellion = untrustworthy, wide sellion = trustworthy), in Study 5 these associations were counterbalanced across participants. After the trust game, there was a brief 20-item pattern recognition distractor task, followed by a surprise memory task wherein participants were asked to discriminate between faces presented during the learning phase vs. those presented during the trust game. The task included 8 randomly selected faces from the learning phase and 8 randomly selected faces from the trust game. At the beginning of the memory task, participants were reminded about the two tasks that they had performed so far in the study: a Learning Behaviors task involving seeing several facial targets alongside a specific behavioral description, and the Economic Trust Game, which involved giving individual facial targets a certain amount of money. Each face was presented individually in a randomized order, and the task was to indicate whether the face was from the Learning Behaviors task or the Economic Trust Game. 5 attention checks were interspersed throughout, with the instruction to press a number (1-5).

A probe of participants' explicit awareness of the sellion-width association immediately followed the face memory task. On screen, faces were shown with wide and narrow sellions, and the sellion feature was explicitly defined as the upper part of the nose bridge via a visual illustration. Participants were asked to indicate whether they used the sellion feature while making payments during the economic trust game. Following this probe, they were debriefed about the study's aims and demographic data was collected.

Results and Discussion

Before conducting analyses, 5 participants were removed for failing attention checks during the trust game, and additional 3 participants were removed for failing attention checks during the face memory task. Another 4 participants were removed for triggering error warnings for excessive speed (< 500 ms) on the majority of trials during the learning phase (indicating that they were not closely reading the behaviors associated with each face).

As in Study 1, payment in the trust game was submitted to a 2 (Sellion Width: Trustworthy vs Untrustworthy) x 2 (Facial Trustworthiness: Trustworthy vs Untrustworthy) repeated-measures ANOVA. There was a significant effect of facial trustworthiness, $F(1,87) = 17.33$, $p < 0.0001$, $\eta_p^2 = 0.17$, with more payment allocated to faces to trustworthy vs. untrustworthy faces. There was also the predicted main effect of sellion width, $F(1,87) = 31.27$, $p < 0.0001$, $\eta_p^2 = 0.26$, such that more money was given to faces with trustworthy vs. untrustworthy sellions. The interaction was not significant, $F(1,87) = 0.64$, $p = 0.43$, $\eta_p^2 = 0.007$ (see Figure 6). These effects held even when statistically controlling for facial competence and facial dominance (and their interactions with sellion width and facial trustworthiness) (Supplementary Tables 10 and 11).

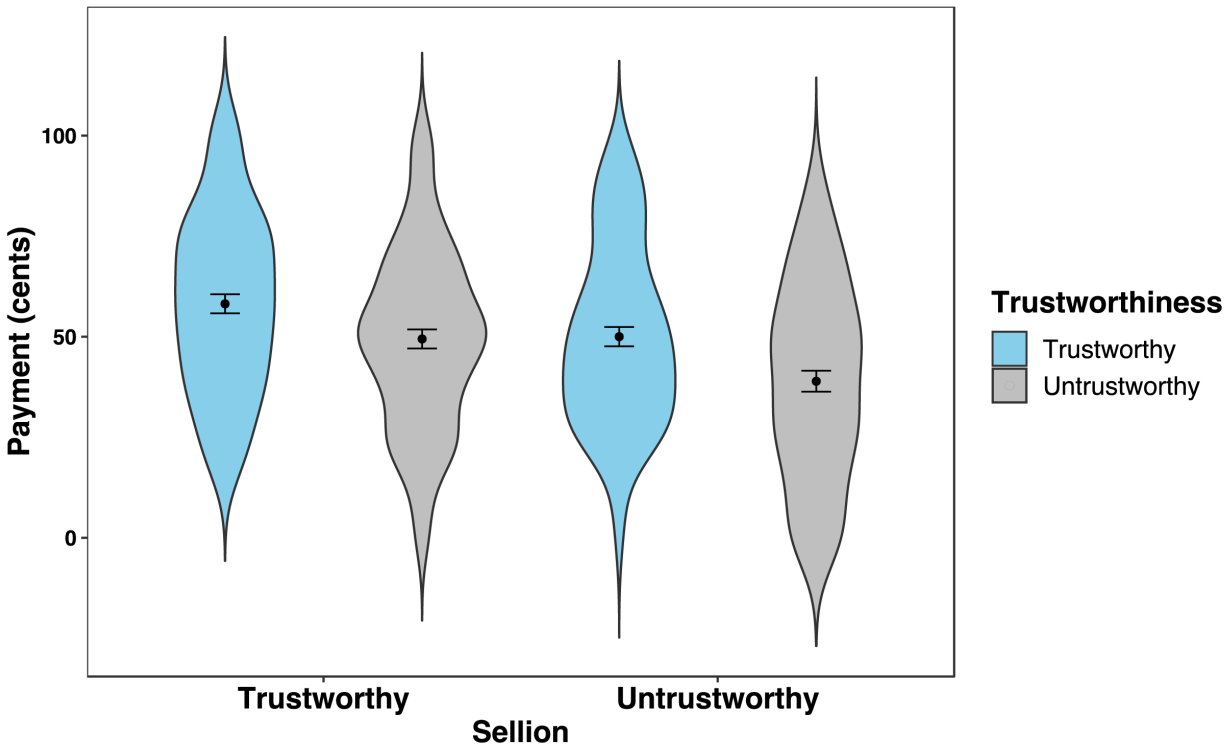

Figure 6. Violin plots representing the probability densities for payment by facial trustworthiness and sellion width for Study 5. Error bars represent the standard error of the mean.

For the surprise face memory task, as is customary in memory recognition tasks we used signal detection analysis to appropriately control for response bias. A hit was defined as correctly identifying a trust-game face as belonging to the trust game, while a miss was defined as incorrectly identifying a trust-game face as belonging to the learning phase. A correct rejection was defined as correctly identifying that a learning-phase face was from the learning phase, while a false alarm was defined as incorrectly identifying that a learning-phase face was from the trust game. Indeed, a one-sample *t*-test confirmed that participants successfully discriminated between faces originating from the two tasks, with a high level of discriminability (*d*') that was significantly more positive ($M = 1.01$, $SE = 0.10$) than zero, one-sample $t(87) = 10.34$, $p < 0.0001$, $d = 1.10$. This result suggests that the learning effects found throughout our studies was due to generalization to new faces that shared the learned sellion feature (rather than confusing

the identities of the faces in the two tasks).

To test awareness of the sellion-width association, we examined responses to the post-task probe. A small minority of participants (18/88 participants, or 20.45%) indicated that they used the sellion width feature to inform their payment decisions. To determine whether the overall effects were driven by this subsample of participants reporting explicit awareness, the data were submitted to a 2 (Sellion Width: Trustworthy vs. Untrustworthy) x 2 (Facial Trustworthiness: Trustworthy vs. Untrustworthy) x 2 (Probe: Aware vs. Unaware) mixed-model ANOVA. There was no main effect of Probe, $F(1,87) = 2.13$, $p = 0.15$, $\eta_p^2 = 0.024$, and critically, there were no significant interactions between Probe and Sellion Width, $F(1,87) = 1.38$, $p = 0.24$, $\eta_p^2 = 0.016$, or Probe and Facial Trustworthiness, $F(1,87) = 0.17$, $p = 0.69$, $\eta_p^2 = 0.002$; and the three-way interaction was also not significant, $F(1,87) = 0.83$, $p = 0.36$, $\eta_p^2 = 0.01$. We additionally reran our analyses after excluding the 18 participants who reported explicit awareness of the sellion-width association, which had a negligible impact on the results. Specifically, with these participants excluded, the effect of facial trustworthiness, $F(1,69) = 13.42$, $p < 0.0001$, $\eta_p^2 = 0.16$, as well as the predicted main effect of sellion, $F(1,69) = 18.26$, $p < 0.0001$, $\eta_p^2 = 0.21$, were both strong and significant. There was no significant interaction, $F(1,69) = 0.90$, $p = 0.77$, $\eta_p^2 = 0.001$. Thus, the vast majority of participants lacked awareness of the sellion-width association, and critically, the effects did not depend on this awareness.

Overall, in Study 5, we replicated Studies 1A/1B's effects and ruled out two potential concerns. Namely, we cast doubt on the possibility that participants confused faces across the learning and evaluation phases and thus did not exhibit genuine generalization. We also demonstrate that, by and large, participants do not have explicit awareness of the mapping between features and behaviors, and moreover, none of our results depend on this awareness.

META-ANALYSIS

To characterize the overall strength of the newly learned facial stereotype effect, we meta-analyzed Studies 1A, 1B, 4, and 5 using fixed effects, with effect size weighted by sample size (Goh, Hall, & Rosenthal, 2016). We did not include the mouse-tracking study (Study 3) as the predicted empirical pattern in that study (Sellion Width × Facial Trustworthiness interaction on mouse-trajectory attraction) was qualitatively different than the other studies (effect of Sellion Width on evaluations). We analyzed effect sizes for Sellion Width, Facial Trustworthiness, and the Sellion Width × Facial Trustworthiness interaction across the four studies. Effect sizes were converted to Cohen's *d*. The meta-analytic effect of Sellion Width was significant, with a conventionally medium effect size, mean $d = 0.54$, $Z = 7.91$, $p < .0001$. The meta-analytic effect of Facial Trustworthiness was also significant, with a conventionally small-to-medium effect size, mean $d = 0.39$, $Z = 5.76$, $p < .0001$. The meta-analytic effect of the Sellion Width × Facial Trustworthiness interaction was not significant, $d = 0.10$, $Z = 1.45$, $p = 0.15$. Thus, although the interaction of both cues on trust-related evaluations was observed in one study (Study 1A), across studies there is insufficient evidence to conclude that the cues have an interactive or additive role in evaluations. Instead, there is strong evidence across studies to conclude that both long-established facial trustworthiness and, more critically, the newly learned sellion feature is able to affect trust-related evaluations.

GENERAL DISCUSSION

Overall, we demonstrate that newly learned associations between an arbitrary facial feature and valenced behaviors resulted in the creation of a novel facial stereotype associated with sellion width, which exerted an equal – if not stronger – effect to the long-established trustworthiness of the face on evaluations. These results arose in payments allocated in an

economic trust game and in the choice of hiring a financial advisor, demonstrating how these newly learned facial stereotype associations affect consequential decision-making after only a brief learning period (Studies 1 and 2). Moreover, these associations had an early impact on the evaluative process on par with facial trustworthiness, as measured by mouse-tracking (Study 3), and this impact was automatic and affected even implicit evaluations indicated by evaluative priming (Study 4). Finally, we ruled out the possibility that the effects were driven by explicit knowledge of the sellion-width association and confirm that the facial stereotypes generalized to new faces that share the learned feature (Study 5). In each study, participants were abstracting the underlying statistical associations between behaviors and the sellion feature and applied this information to evaluations of new faces that also contained those features, showing that the learning generalized to new exemplars at a categorical level, thereby creating a genuine facial stereotype.

These results demonstrate that associative learning has a role in shaping face evaluations, and newly learned feature-behavior associations can be integrated with features of the face that typically denote trustworthiness, effectively resulting in the creation of a novel stereotype based on the learned association between behaviors and specific facial features. That these associations can be learned and generalized to new faces containing those features adds to a growing literature demonstrating that trait-based evaluations of faces are not fixed, and that factors such as learning and experience in the social environment play a role in shaping such evaluations (Dotsch et al., 2016; Hehman et al., 2017; Sofer et al., 2017; Stolier et al., 2018; Stolier et al., 2020). Even if the ability to make trait inferences from faces is present from an early age (Cogsdill et al., 2014), our results suggest that these representations are malleable and update based on newly learned behavioral associations that are mapped onto specific face features.

It is important to note that the learning paradigm involved participants encoding a probabilistic covariation between features and behaviors using third-party information. The aim was to mimic covariations as might be learned over many years from media representations, other people's descriptions, and the broader social context, not necessarily any accurate covariations directly learned from the actual behavior of targets themselves. We do not believe perceivers implicitly learn trustworthiness by linking certain facial features to targets' actual behavior (as there is little correspondence with actual behavior; Wilson & Rule, 2015), but rather that facial stereotypes about people with certain facial features are implicitly communicated to perceivers, and in turn learned and automatically deployed. Regardless, either form of statistical covariation learned by perceivers could give rise to the effects obtained here.

The current results have important theoretical implications for face impressions research. It has been well established that learning positive or negative information about specific individuals impacts their subsequent evaluation (Bliss-Moreau, Barrett, & Wright, 2008; Falvello, Vinson, Ferrari, & Todorov, 2015), and the effects of this learned information can transfer to faces that are perceptually similar (Verosky & Todorov, 2010, 2013). The current work expands upon these findings by demonstrating a more fundamental form of learning and updating beyond individual exemplars, such that new associations between specific facial features themselves and particular traits are generated, leading to the creation of a novel facial stereotype. These facial stereotypes evaluations generalized on a category level to new, subsequently encountered faces with associated features. The rapid acquisition of these stereotypes and their significant effect on evaluations suggests a fundamental malleability of our trait representations based on the abstraction of the co-occurrence between facial features and trait-related behavior. Further study of these statistical learning mechanisms could be beneficial

in understanding the origins of facial stereotypes and how they might be maintained or dismantled based on receiving new information. Specifically, understanding the conditions under which this learning is optimal could be potentially useful in mitigating the more deleterious impacts of facial stereotypes. Further research could also examine other potential boundary conditions, such as a face's location in featural face space (Valentine, 1991). For example, it is possible that more extreme divergences from the average in featural face space could facilitate learning for more extreme behaviors, and so similar facial stereotypes may be more difficult to acquire if the facial features being learned are less extreme. There may be other moderators that could govern the acquisition of these stereotypes, such as exposure level, that could facilitate these learning effects, and future research should endeavor to explore these.

Automatic trait inferences from faces have long been thought to yield functional benefits (Berry & McArthur, 1986; Oosterhof & Todorov, 2008; Little, Jones, & DeBruine, 2011), and face evaluations such as trustworthiness are highly consequential yet have little accurate basis (Eberhardt, Goff, Purdie, & Davies, 2004; Little et al., 2007; Wilson & Rule, 2015). Given these potentially severe consequences, a recent study used a similar statistical learning paradigm as a form of bias intervention (Chua and Freeman, 2020), resulting in the reduction or elimination of biases associated with facial stereotypes. The current work further demonstrates the flexibility of this implicit learning system by extending the results to the creation of entirely new feature associations that exert an influence on trustworthiness evaluations on par with existing facial trustworthiness cues.

It is an open question the extent to which this kind of brief learning paradigm used in the present studies can affect face evaluations in the long-term. We have acquired over a lifetime of representations of faces that have been associated with trustworthy or untrustworthy evaluations,

and as noted previously, our face evaluation abilities are present from a very young age (Cogsdill et al., 2014), so it might seem a daunting task to impact these evaluations in the long-term. However, it is important to note that the learning that took place in these studies was minimal, involving a few minutes of reading short behavioral sentences associated with a relatively small set of faces. Even given this minimal learning, trait evaluations were significantly impacted at explicit and implicit levels across several measures, even influencing hiring decisions. There is evidence that longer, more intensive interventions can result in long-term changes in automatized evaluations and reduced bias (Devine, Forscher, Austin, & Cox, 2012; McNulty, Olson, Jones, & Acosta, 2017), so a more in-depth training might result in shifts in trait evaluations that persist over time. Additionally, a more naturalistic presentation of the behaviors (e.g., in a newspaper headline format or by emulating social media posts) could examine how similarly skewed presentation of information might impact evaluations of trustworthiness. Future studies should also examine these learning effects generalize to other trait dimensions (e.g., competence or dominance). Our theoretical account certainly views any trait dimension to be a candidate for the effects reported here, so long as there is a co-occurrence between facial features and the trait's related behaviors, but this generality must be confirmed with further investigation.

In short, a brief period of learning can result in shifts in both explicit and implicit trait evaluation of faces – shifts comparable to the "intrinsic" facial features that typically convey those traits. These results suggest that our evaluations of faces are not fixed but rather highly malleable, rapidly adapting to newly learned associations from the social world and giving rise to facial stereotypes.

ACKNOWLEDGEMENTS

This work was supported by the NSF BCS-1654731. We thank Michael Berkebile and Maryam Beshar for their assistance with the studies.

REFERENCES

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion, 6*(2), 269-278. doi:10.1037/1528-3542.6.2.269

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior, 10*(1), 122-142.

Berry, D. S., & McArthur, L. Z. (1986). Perceiving character in faces: The impact of age-related craniofacial changes on social perception. *Psychological Bulletin, .100*(1), pp. doi:10.1037/0033-2909.100.1.3 3526376

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* 187-194.

Bliss-Moreau, E., Barrett, L. F., & Wright, C. I. (2008). Individual differences in learning the affective value of others under minimal conditions. *Emotion, 8*(4), 479.

Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2013). The modular nature of trustworthiness detection. *J Exp Psychol Gen, 142*(1), 143-150. doi:10.1037/a0028930

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cogn Psychol, 61*(2), 87-105. doi:10.1016/j.cogpsych.2010.03.001

Chua, K.-W., & Freeman, J. B. (2020). Facial Stereotype Bias Is Mitigated by Training. *Social Psychological and Personality Science*, 1948550620972550.

Cogsdill, E. J., Todorov, A. T., Spelke, E. S., & Banaji, M. R. (2014). Inferring character from faces: a developmental study. *Psychol Sci, 25*(5), 1132-1139. doi:10.1177/0956797614523297

De Neys, W., Hopfensitz, A., & Bonnefon, J. F. (2017). Split-Second Trustworthiness Detection From Faces in an Economic Game. *Experimental Psychology, 64*(4), 231-239. doi:10.1027/1618-3169/a000367

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *J Exp Soc Psychol, 48*(6), 1267-1278. doi:10.1016/j.jesp.2012.06.003

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

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychol Sci, 17*(5), 383-386. doi:10.1111/j.1467-9280.2006.01716.x

Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing black: race, crime, and visual processing. *J Pers Soc Psychol, 87*(6), 876-893. doi:10.1037/0022-3514.87.6.876

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *J Cogn Neurosci, 19*(9), 1508-1519. doi:10.1162/jocn.2007.19.9.1508

Falvello, V., Vinson, M., Ferrari, C., & Todorov, A. (2015). The robustness of learning about the trustworthiness of other people. *Social Cognition, 33*(5), 368-386.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods, 39*(2), 175-191.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion, .15*(2), pp. doi:10.1080/0269993004200024

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*(2), 229-238. doi:10.1037/0022-3514.50.2.229

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77-83.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878.

Freeman, J. B. (2014). Abrupt category shifts during real-time person perception. *Psychon Bull Rev, 21*(1), 85-92. doi:10.3758/s13423-013-0470-8

Freeman, J. B. (2018). Doing psychological science by hand. *Curr Dir Psychol Sci, 27*(5), 315-323. doi:10.1177/0963721417746793

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior research methods, 42*(1), 226-241.

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review, .118*(2), pp. doi:10.1037/a0022327 21355661

Freeman, J. B., & Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behav Res Methods, 45*(1), 83-97. doi:10.3758/s13428-012-0225-x

Freeman, J. B., & Johnson, K. L. (2016). More Than Meets the Eye: Split-Second Social Perception. *Trends Cogn Sci, 20*(5), 362-374. doi:10.1016/j.tics.2016.03.003

Freeman, J. B., Stolier, R. M., Ingbretsen, Z. A., & Hehman, E. A. (2014). Amygdala responsivity to high-level social information from unseen faces. *J Neurosci, 34*(32), 10573-10581. doi:10.1523/JNEUROSCI.5063-13.2014

Giner-Sorolla, R., Garcia, M. T., & Bargh, J. A. (1999). The automatic evaluation of pictures. *Social Cognition, .17*(1), pp. doi:10.1521/soco.1999.17.1.76

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta‐analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass, 10*(10), 535-549.

Hartigan, J. A., & Hartigan, P. M. (1985). The Dip Test of Unimodality. *The Annals of Statistics, 13*, 70-84.

Hehman, E., Flake, J. K., & Freeman, J. B. (2015). Static and dynamic facial cues differentially affect the consistency of social evaluations. *Personality and Social Psychology Bulletin, 41*(8), 1123-1134.

Hehman, E., Sutherland, C. A. M., Flake, J. K., & Slepian, M. L. (2017). The unique contributions of perceiver and target characteristics in person perception. *J Pers Soc Psychol, 113*(4), 513-529. doi:10.1037/pspa0000090

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.

Jones, B. C., Debruine, L. M., Little, A. C., & Feinberg, D. R. (2007). The valence of experiences with faces influences generalized preferences. *Journal of Evolutionary Psychology, 5*(1), 119-129.

Little, A. C., Burriss, R. P., Jones, B. C., & Roberts, S. C. (2007). Facial appearance affects voting decisions. *Evolution and Human Behavior, 28*(1), 18-27. doi:10.1016/j.evolhumbehav.2006.09.002

Little, A. C., Jones, B. C., & DeBruine, L. M. (2011). Facial attractiveness: evolutionary based research. *Philos Trans R Soc Lond B Biol Sci, 366*(1571), 1638-1659. doi:10.1098/rstb.2010.0404

Martin, D., & Macrae, C. N. (2007). A face with a cue: Exploring the inevitability of person categorization. *European Journal of Social Psychology, .37*(5), pp. doi:10.1002/ejsp.445

McArthur, L. Z., & Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological review, 90*, 215-238.

McNulty, J. K., Olson, M. A., Jones, R. E., & Acosta, L. M. (2017). Automatic associations between one's partner and one's affect as the proximal mechanism of change in relationship satisfaction: Evidence from evaluative conditioning. *Psychological Science, 28*(8), 1031-1040.

Montepare, J. M., & Dobish, H. (2003). The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of Nonverbal Behavior, 27*(4), 237-254. doi:Doi 10.1023/A:1027332800296

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends Cogn Sci, 18*(11), 566-570. doi:10.1016/j.tics.2014.09.007

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proc Natl Acad Sci U S A, 105*(32), 11087-11092. doi:10.1073/pnas.0805664105

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLoS One, 7*(3), e34293. doi:10.1371/journal.pone.0034293

Sofer, C., Dotsch, R., Oikawa, M., Oikawa, H., Wigboldus, D. H. J., & Todorov, A. (2017). For Your Local Eyes Only: Culture-Specific Face Typicality Influences Perceptions of Trustworthiness. *Perception, 46*(8), 914-928. doi:10.1177/0301006617691786

Stolier, R. M., Hehman, E., & Freeman, J. B. (2020). Trait knowledge forms a common structure across social cognition. *Nature Human Behavior*.

Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proc Natl Acad Sci U S A, 115*(37), 9210-9215. doi:10.1073/pnas.1807222115

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623-1626. doi:10.1126/science.1110589

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: determinants, consequences, accuracy, and functional significance. *Annu Rev Psychol, 66*, 519-545. doi:10.1146/annurev-psych-113011-143831

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends Cogn Sci, 12*(12), 455-460. doi:10.1016/j.tics.2008.10.001

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. doi:10.1037/0022-3514.83.5.1051

Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *Journal of Experimental Social Psychology, 39*(6), 549-562. doi:10.1016/s0022-1031(03)00059-3

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A, 43*(2), 161-204.

van 't Wout, M., & Sanfey, A. G. (2008). Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. *Cognition, 108*(3), 796-803. doi:10.1016/j.cognition.2008.07.002

Verosky, S. C., Porter, J., Martinez, J. E., & Todorov, A. (2018). Robust effects of affective person learning on evaluation of faces. *J Pers Soc Psychol, 114*(4), 516-528. doi:10.1037/pspa0000109

Verosky, S. C., & Todorov, A. (2010). Generalization of affective learning about faces to perceptually similar faces. *Psychological Science, 21*(6), 779-785.

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.

Willis, J., & Todorov, A. (2006). First impressions: making up your mind after a 100-ms exposure to a face. *Psychol Sci, 17*(7), 592-598. doi:10.1111/j.1467-9280.2006.01750.x

Wilson, J. P., & Rule, N. O. (2015). Facial Trustworthiness Predicts Extreme Criminal-Sentencing Outcomes. *Psychol Sci, 26*(8), 1325-1331. doi:10.1177/0956797615590992

Zebrowitz, L. A., & Montepare, J. M. (2008). Social Psychological Face Perception: Why Appearance Matters. *Soc Personal Psychol Compass, 2*(3), 1497. doi:10.1111/j.1751-9004.2008.00109.x