*Research Report*

Psychological Science
2016, Vol. 27(2) 282–288

DOI: 10.1177/0956797615594897
pss.sagepub.com

# Internalized Impressions: The Link Between Apparent Facial Trustworthiness and Deceptive Behavior Is Mediated by Targets' Expectations of How They Will Be Judged

**Michael L. Slepian and Daniel R. Ames**
Department of Management, Columbia University

**Abstract**

Researchers have debated whether a person's behavior can be predicted from his or her face. In particular, it is unclear whether people's trustworthiness can be predicted from their facial appearance. In the present study, we implemented conceptual and methodological advances in this area of inquiry, taking a new approach to capturing trustworthy behavior and measuring targets' own self-expectations as a mediator between consensual appearance-based judgments and the trustworthiness of targets' behavior. Using this novel paradigm to capture 900 observations of targets' behavior (as trustworthy or untrustworthy), we found that face-based judgments predicted trustworthiness. We also found that this effect was mediated by targets' expectations of how other people would perceive them and by their intentions to act in accordance with those expectations. These results are consistent with an internalized-impressions account: Targets internalize other people's appearance-based expectations and act in accordance with them, which leads facial-appearance-based judgments to be accurate.

**Keywords**

social cognition, social perception, face perception, open data, open materials

Received 2/21/15; Revision accepted 6/16/15

For decades, scholars have debated whether a person's behavior can be predicted from his or her face. In particular, can judgments of individuals' faces predict their trustworthiness? One possibility (the *essentialist-impressions* account) is that genetic expression leads to both untrustworthy-looking faces and untrustworthy behavior. Such a correspondence would resemble the once-popular but now discredited claims of physiognomy. An alternative possibility (the *misleading-impressions* account) is that although people reliably agree on which faces look untrustworthy or trustworthy (Rule, Krendl, Ivcevic, & Ambady, 2013; Todorov, 2008), those judgments show no predictive validity (Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015) because there is no reliable correspondence with actual trustworthiness (see Todorov & Porter, 2014). We believe that a third possibility exists: A lifetime of being treated as trustworthy or untrustworthy as a result of one's appearance may lead one to internalize these expectations and act in accordance with them, which eventually results in appearance-based accuracy (our *internalized-impressions* account).

Recent research presents a mixed picture: Some work has provided evidence suggestive of accuracy in face-based judgments of trustworthiness (Stirrat & Perrett, 2010), yet other research has found no such accuracy (Rule et al., 2013). We see two reasons why prior work might have obtained mixed results. First, some past scholarship has revolved around single, relatively extreme,

**Corresponding Author:**
Michael L. Slepian, Department of Management, Columbia Business School, Columbia University, Uris Hall, 3022 Broadway, New York, NY 10027
E-mail: michael.slepian@columbia.edu

and heterogeneous behaviors (e.g., targets' criminal history). Instead, we think a relationship between faces and behavior bears testing in a paradigm with three features: (a) an interactive face-to-face context, (b) a constrained set of trust-related behaviors, and (c) multiple observations of potentially trustworthy or untrustworthy behavior. Second, prior studies have not measured psychological mediators between face-based judgments and behavior. We expect that any link between facial appearance and behavior would be mediated by psychological variables, such as targets' expectations.

In this article, we introduce a new paradigm that addresses both of these points. We created a novel research design to capture multiple instances of trust-related behaviors in a face-to-face context. We focused on a single class of behaviors: People (targets) repeatedly chose to make and defend a true or false claim to different counterparts, a false claim (i.e., deceptive, untrustworthy behavior) entailing the chance for private material gains but also imposing costs on the counterpart. Past paradigms have typically placed participants either in computer-mediated interactions or in asocial contexts. We expected that giving targets the option to lie (for potential gain) to a live face-to-face counterpart would elicit meaningful variance in trustworthiness of behavior not captured in prior research. Our paradigm also constrained behavior into a dichotomous choice (to lie or tell the truth) made repeatedly (in 10 independent interactions with different counterparts), yielding a clean and reliable measure of trustworthiness.

In addition, before interactions, we assessed targets' metaperceptions of their own trustworthiness (targets predicted how frequently they would be trusted by their counterparts) and their predictions of how frequently they would act in a trustworthy manner. Measuring these variables allowed us to test a potential route through which judgments of the face might predict behavior. That is, given that people reach high consensus on which faces look trustworthy or untrustworthy, individuals with trustworthy- or untrustworthy-looking faces should have a lifetime of experience of being treated like trustworthy or untrustworthy people. Such experiences would range from the banal (e.g., whether strangers smile at them) to the life-changing (e.g., whether they get particular jobs). We believe that the cumulative effect of such treatment is likely to be powerful, as implied by work on self-fulfilling prophecies (Rosenthal, 1994) and the looking-glass self (Cooley, 1902).

## Method

We first measured participants' apparent facial trustworthiness by having independent judges rate photographs of them. Two days later, participants were told that they would be interacting with other participants and were asked to report how they expected to be judged by their counterparts and how they expected themselves to act. They subsequently interacted as both targets and counterparts in a novel mixed-motive game. As targets, they repeatedly chose whether to behave in a trustworthy manner (i.e., to tell the truth) or in an untrustworthy manner (i.e., to lie) to a series of 10 different counterparts. We predicted that if ratings of facial trustworthiness showed an ability to predict trustworthy behavior, then this link would be mediated by targets' expectations, which would be consistent with an internalized-impressions account.

### *Participants*

Our participant pool consisted of all the M.B.A. students in a particular course; sample size was determined by the number of students who were enrolled in the course and present on the day the study was conducted ($N$ = 118). Ninety-five participants' faces were photographed, but 5 of these participants did not provide self-expectation judgments. Thus, the final sample consisted of 90 students (65.60% male; mean age = 28.10 years, $SD$ = 1.76).

### *Mixed-motive game*

Participants played a two-person game in which each person privately drew a random card (labeled "high" or "low"). In a face-to-face interaction, they then freely chose to claim that the card was "high" or "low," independently of the card drawn, thereby choosing to tell the truth or to lie.

The mixed-motive paradigm was implemented in two testing sessions (accounting for testing session in our analyses did not alter the results). Targets and counterparts were randomly paired within sessions, with no repeat pairings. Each participant was randomly paired with 10 other participants in succession. In each interaction, a given participant served as both a target and a counterpart. As the target, the participant decided whether to tell the truth or to lie to the counterpart; as the counterpart, the participant decided whether to trust the target. After both members of a pair had drawn a card, and independently and privately decided whether to tell the truth or to lie, they then claimed that their cards were either high or low (thereby telling the truth or not). They next spent 2 to 3 min attempting to persuade one another of their trustworthiness. After this persuasion phase, each participant independently and privately judged whether he or she trusted that the target was telling the truth. Once both parties had made their private judgments about one another, both revealed whether they had lied or told the truth, and whether they had trusted their fellow participant.

**Table 1.** Payoff Table for a Single Round of Mixed-Motive Game

| | Partner's trust in participant | | | |
|---|---|---|---|---|
| | Partner (counterpart) trusts participant (target) | | Partner (counterpart) distrusts participant (target) | |
| Participant's trust in partner | Participant is trustworthy (+10 for participant) | Participant is untrustworthy (+20 for participant) | Participant is trustworthy (0 for participant) | Participant is untrustworthy (0 for participant) |
| Participant (counterpart) trusts partner (target) | | | | |
| Partner is trustworthy (+10 for participant) | +20 [+20] | +30 [−10] | +10 [+10] | +10 [+10] |
| Partner is untrustworthy (−20 for participant) | −10 [+30] | 0 [0] | −20 [+20] | −20 [+20] |
| Participant (counterpart) distrusts partner (target) | | | | |
| Partner is trustworthy (0 for participant) | +10 [+10] | +20 [−20] | 0 [0] | 0 [0] |
| Partner is untrustworthy (0 for participant) | +10 [+10] | +20 [−20] | 0 [0] | 0 [0] |

Note: A given participant's payoff for a given round was determined by two components: The participant's score as the counterpart (i.e., whether the participant trusted his or her partner) and the participant's score as a target (i.e., whether the participant's partner trusted the participant). As counterparts, participants earned points for trusting their partners when their partners were trustworthy and lost points for trusting their partners when their partners were untrustworthy. As targets, participants earned points for getting their partners to trust them and gained no points if their partners did not trust them. Each cell contains two values; the number outside the brackets shows the total payoff for the person identified as the participant, and the number inside the brackets shows the total payoff for that participant's partner.

We describe the outcome of the decision to tell the truth or to lie as trustworthy or untrustworthy *behavior*. The only way for targets to earn points on the basis of their behavior was to earn trust. If a target chose to tell the truth and was trusted by his or her counterpart, the target earned a modest payoff in the game (10 points). If a target chose to lie and was trusted by his or her counterpart, the target earned double that payoff (20 points). If a target's counterpart did not trust the target, the target earned nothing (0 points). In game-theory terms, lying was a weakly dominant strategy (see Kohlberg & Mertens, 1986).

We describe the outcome of the decision to trust or distrust a target as a *judgment*. Counterparts' judgments about whether to trust targets also had payoffs. If a target's counterpart correctly trusted a target who told the truth, the counterpart received a modest reward (10 points). Incorrectly trusting a target who lied entailed a significant loss (−20 points). If a counterpart decided to not trust the target, he or she neither earned nor lost points (0 points). Thus, counterparts' payoffs for their trust judgments were contingent on whether a target was telling the truth or lying. In game-theory terms, there was no dominant strategy for judgments (if participants assumed that lies and truths were equally likely but undiagnosable). The three top performers in each session received prizes (a $50 Amazon gift card for the top performer and $25 Amazon gift cards to the second- and third-place performers).

The payoff table shown in Table 1 summarizes participants' payoffs for each possible outcome. In game-theory terms, lying is a weakly dominant strategy; if we assume that each player recognizes that lying is weakly dominant for the other, lying combined with distrusting is the Nash equilibrium (see Kohlberg & Mertens, 1986). However, we did not expect that most interactions would result in mutual lying and distrust. Note that in each of the 10 rounds, each participant was both a target (choosing how to behave) and a counterpart (judging a fellow participant). These choices were made separately, with behavioral decisions made privately before mutual discussion (i.e., behavioral choices were made before the interaction) and judgments made privately after discussion. Thus, the calculation of the payoff matrix should not be taken to suggest that one decision was contingent on the other; they were independent. We designed this paradigm with the expectation that it would produce variance in the frequency with which people would behave in a trustworthy manner (i.e., some people would choose to lie frequently, and others would choose to tell the truth frequently); this variance was critical for testing our predictions.

In this study, we examined whether individuals' behavior (not their judgments of other people) could be predicted from their faces, and thus the focus of our analysis was predicting targets' behavior toward others, not counterparts' judgments of others. Our two central questions were (a) whether the apparent facial trustworthiness of targets (based on ratings from an independent set of judges) predicted how they behaved (i.e., their frequency of telling the truth) in the 10-round game and (b) whether that link was mediated by the targets' expectations reported before the game.

### Measures

Two days before the game, during a video-based exercise occurring in the students' class, photographs were taken of the players. In the photographs, the players assumed a neutral expression. No specific rationale was given for taking photographs other than that it was part of the video-based class exercise. That is, these photographs were taken outside the context of the mixed-motive game; participants were not aware of the game or its rules when the photographs were taken. We recruited independent judges ($n$ = 30 per rating) via Amazon.com's Mechanical Turk. These judges used a 7-point scale to rate each face for trustworthiness (1 = *not at all trustworthy*, 7 = *very trustworthy*; $M$ = 4.127, $SD$ = 0.616, 95% confidence interval, or CI = [3.998, 4.256], $\alpha$ = .887), attractiveness (1 = *not at all attractive*, 7 = *very attractive*; $M$ = 3.264, $SD$ = 0.805, 95% CI = [3.096, 3.433], $\alpha$ = .948), *babyfaceness* (1 = *not at all babyfaced*, 7 = *very babyfaced*; $M$ = 3.332, $SD$ = 0.811, 95% CI = [3.162, 3.502], $\alpha$ = .918), and apparent affect (1 = *appears angry*, 7 = *appears happy*; $M$ = 3.962, $SD$ = 0.830, 95% CI = [3.788, 4.135], $\alpha$ = .952).

During the game session, after the process and payoffs had been described to the players, they predicted how frequently (0%–100%) they would (a) act in a trustworthy manner ($M$ = 52.6%, $SD$ = 31.8, 95% CI = [45.9%, 59.2%]) and (b) be trusted ($M$ = 55.2%, $SD$ = 17.5, 95% CI = [51.6%, 58.9%]).[1] The game yielded 10 observations of each participant's behavior as a target, the focal measure for the current study (coded as 1 = trustworthy, 0 = untrustworthy; $M$ = 62.4%, $SD$ = 35.5, 95% CI = [55.0%, 69.9%]), and 10 observations of each participant's judgment (of other targets) as a counterpart (coded as 1 = trust, 0 = distrust; $M$ = 61.7%, $SD$ = 18.4, 95% CI = [57.8%, 65.5%]).

Given the nature of the student sample, players had varying levels of familiarity with each other before the game, and familiarity could influence behavior. To assess and control for this possibility, we provided the players with a list of their 10 counterparts' names 4 days after the game and asked them to rate how familiar they had been with each counterpart before the game (1 = *not at all*, 2 = *slightly*, 3 = *somewhat*, 4 = *mostly*, 5 = *highly*; $M$ = 2.264, $SD$ = 1.515, 95% CI = 2.164, 2.363]).

## Results

To maximize statistical power, we fitted outcomes to a linear mixed-effects model examining all 900 trust judgments and 900 trust behaviors, controlling for random variance (from targets, counterparts, and round of the mixed-motive game). All analyses were conducted in the R software environment (Version 3.1.1; R Development Core Team, 2014). We used the R package lme4 to implement mixed-effects models (Bates, Maechler, Bolker, & Walker, 2015). In calculating $p$ values, we used the R package lmerTest to run lme4 models through Satterthwaite approximation tests to estimate the degrees of freedom (these estimated degrees of freedom scale the model estimates to best approximate the $F$ distribution, and thus can be fractional and differ slightly across tests; Kuznetsova, Brockhoff, & Christensen 2013). R package confint was used to implement Wald-tests to calculate 95% CIs.

### Perception

Photograph-based trustworthiness judgments predicted how often counterparts chose to trust targets after the live interactions, $b$ = 0.074, 95% CI = [0.022, 0.126], $SE$ = 0.027, $t(79.76)$ = 2.77, $p$ = .007. Counterparts had access to a multitude of cues in the face-to-face interactions, yet their trustworthiness judgments corresponded to independent ratings of the targets' faces. If this effect emerged in the absence of accuracy, our results would fit with the misleading-impressions account noted earlier. However, our internalized-impressions account suggests that trustworthiness judgments could be accurate, and we next tested for such accuracy.

### Accuracy

Photograph-based trustworthiness judgments predicted how often targets actually behaved in a trustworthy manner toward counterparts, $b$ = 0.124, 95% CI = [0.007, 0.242], $SE$ = 0.060, $t(87.99)$ = 2.077, $p$ = .041. This finding is consistent with the notion that facial trustworthiness predicts trustworthy behavior.

### Other predictors

Before turning to our central prediction (concerning how targets' expectations might mediate the link between their facial trustworthiness and trustworthy behavior), we considered a number of other possible predictors and alternative explanations. Accuracy of trustworthiness judgments

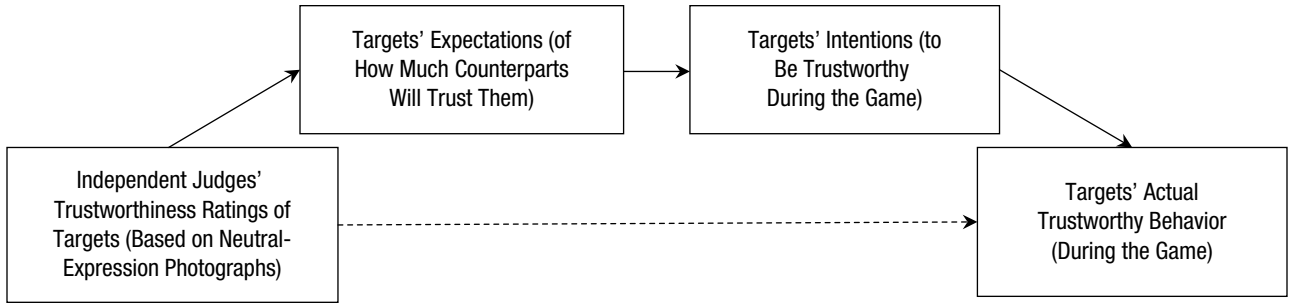


**Fig. 1.** Mediation model of the predicted mediation between targets' facial trustworthiness (as judged by the independent raters) and the targets' behavior, as mediated by the targets' expectations of how they would be judged and how they would act.

might derive from other features of targets' neutral-expression faces (e.g., emotional resemblances; Hehman, Flake, & Freeman, 2015; Sacco & Hugenberg, 2009; Zebrowitz, 2011). Even though the photographs were taken outside the context of and before the game, perhaps targets still somehow conveyed their trustworthy intentions (e.g., by smiling slightly). Our findings were inconsistent with this suggestion: Ratings of attractiveness corresponded with counterparts' trustworthiness judgments, $b = 0.042$, 95% CI = [0.001, 0.083], $SE = 0.021$, $t(81.62) = 1.996$, $p = .049$, but other variables did not—apparent affect: $b = 0.007$, 95% CI = [–0.034, 0.047], $SE = 0.021$, $t(80.89) = 0.331$, $p = .741$; babyfaceness: $b = 0.025$, 95% CI = [–0.018, 0.067], $SE = 0.022$, $t(85.17) = 1.147$, $p = .255$; target's gender (0 = male, 1 = female), $b = 0.060$, 95% CI = [–0.009, 0.129], $SE = 0.035$, $t(80.95) = 1.712$, $p = .091$.

Ratings of babyfaceness predicted trustworthy behavior, $b = 0.103$, 95% CI = [0.014, 0.192], $SE = 0.045$, $t(87.99) = 2.272$, $p = .026$, but other variables did not—attractiveness: $b = 0.022$, 95% CI = [–0.070, 0.114], $SE = 0.047$, $t(87.99) = 0.477$, $p = .635$; apparent affect: $b = 0.065$, 95% CI = [–0.023, 0.153], $SE = 0.045$, $t(87.99) = 1.44$, $p = .153$; target's gender: $b = 0.105$, 95% CI = [–0.048, 0.259], $SE = 0.078$, $t(87.99) = 1.347$, $p = .182$.

It is also possible that the level of familiarity between players influenced the trustworthiness of their behavior. The greater targets' familiarity with their counterparts, the more likely the targets were to behave in a trustworthy manner toward those counterparts, $b = 0.041$, 95% CI = [0.025, 0.057], $SE = 0.008$, $t(835.08) = 5.027$, $p < .001$. Critically, when we accounted for targets' familiarity with their counterparts, greater perceived trustworthiness, as judged from the target's face by the independent raters, was still associated with greater trustworthy behavior, $b = 0.123$, 95% CI = [0.007, 0.240], $SE = 0.060$, $t(88.02) = 2.070$, $p = .041$.

## *Mediation by self-expectations*

We next turned to the mediation prediction implied by the internalized-impressions account. We expected that the link between face-based judgments and behavior would be mediated by the targets' expectations of how they would be judged and how they would act (Fig. 1). Results of regression analyses were consistent with this prediction, revealing that photograph-based trustworthiness judgments predicted targets' self-expectations of how often they would be trusted, $b = 0.073$, $SE = 0.029$, 95% CI = [0.015, 0.131], $t(88) = 2.500$, $p = .014$ (i.e., the targets anticipated other people's naive expectations). These expectations about being trusted, in turn, predicted how often targets intended to act in a trustworthy manner, $b = 0.685$, $SE = 0.180$, 95% CI = [0.328, 1.042], $t(88) = 3.809$, $p < .001$ (i.e., targets internalized these expectations and intended to act consistently with them). Targets' intentions of acting in a trustworthy manner, in turn, predicted the trustworthiness of their actual behavior, $b = 0.805$, $SE = 0.082$, 95% CI = [0.641, 0.968], $t(88) = 9.785$, $p < .0001$ (see Table 2 for zero-order correlations of these variables).[2] A formal bootstrapped mediation analysis (5,000 iterations), in which attractiveness, apparent affect, babyfaceness, and target's gender were entered as covariates, confirmed this mediational path, mean indirect effect = .0512, $SE = .0349$, 95% CI = [.0062, .1548]; excluding the covariates did not alter statistical significance, mean indirect effect = .0363, $SE = 0.0195$, 95% CI = [.0095, .0915].

**Table 2.** Zero-Order Correlations Between the Main Variables in the Mediation Model

| Variable | Targets' facial trustworthiness | Targets' expectations | Targets' intentions |
|---|---|---|---|
| Targets' expectations | .26 | — | — |
| Targets' intentions | .27 | .38 | — |
| Targets' behavior | .22 | .21 | .72 |

Note: All correlations are significant, $p \leq .05$.

## Discussion

In a novel paradigm featuring trusting behavior in face-to-face interactions, trustworthiness ratings of targets (based on neutral-expression photographs) corresponded with targets' behavioral trustworthiness. We found that targets seemed to have an awareness of how people would judge them, and they internalized these expectations and behaved in accordance with them. Such internalized impressions are similar to self-fulfilling prophecies (Rosenthal, 1994), although we suggest that the effects of the internalized impressions are somewhat broader and more cumulative. Much work on self-fulfilling prophecies has focused on a single context, such as whether a teacher's expectations (e.g., that a student is particularly intelligent) can bring about outcomes consistent with those expectations (e.g., improvement in the student's performance; Rosenthal, 1994). We believe that these effects can play out over longer periods as well, as implied by work on the looking-glass self (Cooley, 1902). Our participants' (somewhat accurate) expectations of how other people would judge them in a particular game corresponded to strangers' ratings of photographs of participants' faces, which suggests that these accurate metaperceptions may be derived from a range of contexts across a lifetime of treatment.

Participants' behavior seemed to live up, or down, to how they expected to be judged. Those participants who thought they would be trusted were more likely to be trustworthy, and those who thought they would be distrusted were more likely to be untrustworthy. This finding is inconsistent with an opportunistic-deception account (Olekalns & Smith, 2009), which implies that people who expect to be trusted would be likely to exploit that trust rather than comply with it.

An essentialist-impressions account would not necessarily imply our mediation results. Such an account emphasizes the genetic, inherent correspondence between facial features and behavior rather than the role of targets' expectations posited by the internalized-impressions perspective. It might be possible to further discriminate between essentialist-impressions and internalized-impressions accounts by testing for the causal order that we posit (e.g., manipulating targets' expectations about being trusted, and examining their behavioral trustworthiness).

Features of our new paradigm may have allowed us to observe predictive accuracy of face-based judgments that was not apparent in past work (e.g., Rule et al., 2013). In this study, targets repeatedly confronted a basic but constrained question: "Do I act untrustworthy toward this person for a better chance to win this game?" Placing targets in identical, constrained, highly social contexts, with repeated observations, is likely to increase the robustness of any judgment-behavior link, but that constraint also limits generalizability. The type of photograph-based judgments we used might not predict cheating on a test, for example, but might predict behavior in other mixed-motive social contexts.

In sum, the consensus that people achieve in rating targets' faces corresponds to how people interact with those targets. Targets are aware of the expectations implicit in that consensus. We propose that targets come to internalize such expectations, acting in accordance with them, and thus those initial judgments, over time, become accurate judgments.

### Author Contributions

Both authors contributed to designing the study, analyzing the data, and writing the manuscript.

### Acknowledgments

We thank Patrick Bergemann, Amie Blocker, Ashli Carter, Jaee Cho, Jinseok Chun, Drew Jacoby-Senghor, Alice Lee, Ashley Martin, Pilar Opazo, Anastasia Usova, Stacey Sasaki, and Abbie Wazlawek for their assistance in the current work. We also thank the Social Cognitive & Neural Sciences Lab at New York University for helpful feedback on this work, and Rodney Atkins and Nicholas Rule for helpful comments on an early version of the manuscript.

### Declaration of Conflicting Interests

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

### Open Practices

All data and materials have been made publicly available via Open Science Framework. Data can be accessed at https://osf.io/kfdca/, and materials can be accessed at https://osf.io/bnqd5. The complete Open Practices Disclosure for this article can be found at http://pss.sagepub.com/content/by/supplemental-data. This article has received badges for Open Data and Open Materials. More information about the Open Practices badges can be found at https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/ and http://pss.sagepub.com/content/25/1/3.full.

### Notes

1. We also measured participants' predictions of how much they would trust their fellow participants, how often their fellow participants would act trustworthy, and how accurately they would judge their fellow participants. However, these judgments would be made by the participant in the role of counterpart (rather than target) and are thus outside the scope of the current investigation (which focuses on targets' behaviors, not counterparts' judgments). We also asked participants, in their role as targets, to predict how accurately they would be judged by their counterparts, yet this measure of participants' perceived transparency does not distinguish between behavior as trustworthy or untrustworthy and is thus also outside the scope of the current investigation.

2. Targets' expectations predicted their behavior. Thus, one might wonder whether individual differences in meta-accuracy (i.e., the *difference* between the amount of trust participants expected to receive and the amount that they actually received) would predict behavior. However, meta-accuracy did not correspond to how often targets told the truth, $b = -0.120$, $SE = 0.185$, 95% CI = [–0.487, 0.248], $t(88) = -0.647$, $p = .519$ (nor did the absolute value of the difference, $b = 0.103$, $SE = 0.265$, 95% CI = [–0.424, 0.629], $t(88) = 0.387$, $p = .700$).

## References


Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using 'Eigen' and S4 (Version 1.1-7) [Computer software]. Retrieved from http://cran.r-project.org/package=lme4

Cooley, C. H. (1902). *Human nature and the social order*. New York, NY: Charles Scribner's Sons.

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

Kohlberg, E., & Mertens, J.-F. (1986). On the strategic stability of equilibria. *Econometrica*, *54*, 1003–1037.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2013). lmerTest: Tests in linear mixed effects models (Version 2.0-11) [Computer software]. Retrieved from http://cran.r-proj ect.org/web/packages/lmerTest/

Olekalns, M., & Smith, P. L. (2009). Mutually dependent: Power, trust, affect and the use of deception in negotiation. *Journal of Business Ethics*, *85*, 347–365.

R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rosenthal, R. (1994). Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science*, *3*, 176–179.

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology*, *104*, 409–426.

Sacco, D. F., & Hugenberg, K. (2009). The look of fear and anger: Facial maturity modulates recognition of fearful and angry expressions. *Emotion*, *9*, 39–49.

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

Todorov, A. (2008). Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Annals of the New York Academy of Sciences*, *1124*, 208–224.

Todorov, A., Olivola, C., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*, 519–545.

Todorov, A., & Porter, J. M. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological Science*, *25*, 1404–1417.

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