# Accuracy and Consensus in Judgments of Trustworthiness From Faces: Behavioral and Neural Correlates

Nicholas O. Rule
University of Toronto

Anne C. Krendl
University of Indiana–Bloomington

Zorana Ivcevic
Yale University

Nalini Ambady
Stanford University

Perceivers' inferences about individuals based on their faces often show high interrater consensus and can even accurately predict behavior in some domains. Here we investigated the consensus and accuracy of judgments of trustworthiness. In Study 1, we showed that the type of photo judged makes a significant difference for whether an individual is judged as trustworthy. In Study 2, we found that inferences of trustworthiness made from the faces of corporate criminals did not differ from inferences made from the faces of noncriminal executives. In Study 3, we found that judgments of trustworthiness did not differ between the faces of military criminals and the faces of military heroes. In Study 4, we tempted undergraduates to cheat on a test. Although we found that judgments of intelligence from the students' faces were related to students' scores on the test and that judgments of students' extraversion were correlated with self-reported extraversion, there was no relationship between judgments of trustworthiness from the students' faces and students' cheating behavior. Finally, in Study 5, we examined the neural correlates of the accuracy of judgments of trustworthiness from faces. Replicating previous research, we found that perceptions of trustworthiness from the faces in Study 4 corresponded to participants' amygdala response. However, we found no relationship between the amygdala response and the targets' actual cheating behavior. These data suggest that judgments of trustworthiness may not be accurate but, rather, reflect subjective impressions for which people show high agreement.

*Keywords:* trustworthiness, accuracy, cognitive neuroscience, face perception, criminality

Lay wisdom warns us not to "judge a book by its cover." Yet a growing body of research in social psychology provides evidence of instances in which our first impressions can provide reliable cues to others' thoughts and behaviors (Ambady, Bernieri, & Richeson, 2000; Zebrowitz, 1997). The current work aims to test the validity of one judgment that has been widely studied in the scientific literature and is also of great ecological importance: impressions of individuals' trustworthiness.

Accurately assessing whether or not someone is trustworthy is a highly valued social skill. The extant research has extensively examined perceptions of trustworthiness. For instance, researchers have investigated how we form consensual judgments of trustworthiness (e.g., Krumhuber et al., 2007; Rule, Ambady, & Adams, 2009; Todorov, Pakrashi, & Oosterhof, 2009; Zebrowitz, Voinescu, & Collins, 1996) and how perceptions of trustworthiness can predict individuals' life outcomes (Rule et al., 2010; Zebrowitz et al., 1996).

Perhaps the most thorough exploration of perceived trustworthiness has come from work in social neuroscience that indicates the importance of the amygdala in judgments of trustworthiness. Adolphs and colleagues have shown that damage and dysfunction in the amygdala renders perceivers unable to judge others' trustworthiness, where the criterion for trustworthiness is consensual judgments made by others (Adolphs, Baron-Cohen, & Tranel, 2002; Adolphs, Sears, & Piven, 2001; Adolphs, Tranel, & Damasio, 1998). A number of other researchers have also found that the human amygdala responds to the perceived trustworthiness of

This article was published Online First December 31, 2012.

Nicholas O. Rule, Department of Psychology, University of Toronto, Toronto, Ontario, Canada; Anne C. Krendl, Department of Psychology, Indiana University–Bloomington; Zorana Ivcevic, Department of Psychology, Yale University; Nalini Ambady, Department of Psychology, Stanford University.

This work was supported in part by Natural Sciences and Engineering Research Council of Canada Grant 419593 to Nicholas O. Rule. We would like to thank Josh Lang, Nadia Al-Dajani, David Pham, Jimmy Zuniga, Gali Peleg, Julia Kuelzow, and K. C. Hallett for their assistance with data collection.

Correspondence concerning this article should be addressed to Nicholas O. Rule, Department of Psychology, University of Toronto, 100 St. George Street, Toronto, ON M5S 3G3, or to Nalini Ambady, Department of Psychology, Stanford University, Jordan Hall, Building 420, Stanford, CA 94305. E-mail: rule@psych.utoronto.ca or nambady@stanford.edu

Journal of Personality and Social Psychology, 2013, Vol. 104, No. 3, 409–426
 0022-3514/13/$12.00 DOI: 10.1037/a0031050

faces (e.g., Winston, Strange, O'Doherty, & Dolan, 2002), among other social traits (e.g., Todorov, Baron, & Oosterhof, 2008). The amygdala has been associated with key social judgments such as general evaluations of valence (Oosterhof & Todorov, 2008; Todorov, Said, Engell, & Oosterhof, 2008), evaluations of motivational cues (Cunningham, Van Bavel, & Johnsen, 2008), and arousal (Rule et al., 2011), across numerous studies. It is clear that the amygdala is an important part of the social brain (e.g., Amodio & Frith, 2006), with a special role in the evaluation of other people (Schiller, Freeman, Mitchell, Uleman, & Phelps, 2009). Although agreement among people regarding perceptions of trustworthiness has received considerable attention, an open question remains as to whether these perceptions relate to actual trustworthy behavior. Simply put, are the people perceived to be trustworthy based on their faces actually trustworthy? The current experiments were designed to investigate this question and to clarify the amygdala response to trustworthiness.

A few studies have examined aspects of the predictive validity of judgments of trustworthiness by focusing on specific behaviors assumed to tap trustworthiness as an underlying trait. One study found that participants who were more willing to engage in experiments involving deception were perceived to be less trustworthy (Bond, Berry, & Omar, 1994). Another study found that the faces of America's Most Wanted criminals were rated as less trustworthy than the recipients of great honors, such as the Nobel peace prize (Porter, England, Juodis, ten Brinke, & Wilson, 2008); and a third study showed that violent criminals could be differentiated from nonviolent criminals based on photos of their faces (Stillman, Maner, & Baumeister, 2010). In contrast, Zebrowitz et al. (1996) examined the relationship between perceptions of honesty from photos of individuals' faces taken across the lifespan and found that these judgments did not correspond to clinicians' assessments of their actual honesty based on personality measures. More recently, two studies found that perceptions of trustworthiness were related to whether individuals behaved cooperatively in economic games (Stirrat & Perrett, 2010; Verplaetse, Vanneste, & Braeckman, 2007). These judgments appear to be based primarily on perceptions of aggression, which are strongly negatively correlated with trustworthiness and are communicated by many of the same facial cues (Carré & McCormick, 2008; Carré, McCormick, & Mondloch, 2009). Thus, across a wide range of behaviors (criminal acts, deception, selfish aggression) people who behave deviantly are assumed to be untrustworthy, though the evidence for whether this is reflected in facial appearance is somewhat mixed.

One area of closely related work that has been explored in much detail is the ability to detect deception. Many studies have investigated the accuracy of detecting deception and the cues that perceivers might use to make these judgments (see Bond & DePaulo, 2006, 2008; DePaulo et al., 2003, for reviews). This work has tested the ability of a wide range of individuals, ranging from undergraduate psychology students to law enforcement agents, to reliably and accurately decode whether others are lying versus telling the truth. The general conclusion drawn from this research is that the average person's ability to detect deception is only moderately above chance (Bond & DePaulo, 2006, 2008), mainly because people typically attend to misleading cues in judging others' honesty and ignore the cues that are truly revealing (DePaulo et al., 2003). Indeed, although perceivers tend to show high consensus in whom they believe to be honest and dishonest, this tends not to be related to targets' actual behavior (DePaulo & Rosenthal, 1979).

Although consensus is often not correlated with accuracy in deception detection, consensus and accuracy are correlated in many other areas. Two traits that seem to be accurately perceived are extraversion and intelligence. Perceivers' judgments about the extraversion of others have shown significant correlations with the individuals' self-reports of extraversion and reports by close acquaintances (e.g., Borkenau & Liebler, 1993). Indeed, judgments of others' extraversion are reliable across a variety of different media, including videos, audio clips, still photos, and Facebook pages (Borkenau & Liebler, 1992; Naumann, Vazire, Rentfrow, & Gosling, 2009; Penton-Voak, Pound, Little, & Perrett, 2006; Weisbuch, Ivcevic, & Ambady, 2009). Similar results have been found for intelligence. Perceivers' impressions of intelligence correlate with targets' self-reports and the perceptions of acquaintances (Borkenau & Liebler, 1993). Perhaps most interesting, these impressions also significantly correlate with the targets' actual intelligence, as measured by intelligence quotient (IQ) assessments (Borkenau & Liebler, 1993; Murphy, 2007; Murphy, Hall, & Colvin, 2003; Zebrowitz, Hall, Murphy, & Rhodes, 2002; Zebrowitz & Rhodes, 2004). Thus, not only do people show strong agreement in their assessments of extraversion and intelligence, but these assessments can also be validated through various measures.

Given the practical importance of perceptions of trustworthiness, the present investigation sought to test whether impressions of trustworthiness from faces relate to actual behavior across several ecologically valid domains. Operationalizations of trustworthiness have been mostly implied and highly diverse (e.g., Bond et al., 1994; Stirrat & Perrett, 2010; Verplaetse et al., 2007; Zebrowitz et al., 1996), with one model suggesting that perceptions of trustworthiness may be conceptually and physically based on overgeneralizations of perceived emotional facial expressions such that facial features resembling happiness promote inferences of targets as trustworthy and facial features resembling anger promote inferences of targets as untrustworthy (Oosterhof & Todorov, 2008; see also Todorov, Baron, et al., 2008; Zebrowitz & Montepare, 2008). Particularly unclear is whether trustworthiness constitutes a stable trait or an ephemeral behavior, including the type and quantity of behaviors needed to credential someone as trustworthy or untrustworthy. Absent a clearly demarcated universal definition of trait trustworthiness, previous studies have relied on associating perceptions of trustworthiness with discrete behaviors that are presumed to tap trait trustworthiness as a latent construct. To address the relatively intangible nature of trustworthiness as a trait, in the present work we attempted to test a diverse array of behaviors that are generally perceived in the literature to be (un)trustworthy across several experiments, thereby converging on the underlying trait of trustworthiness from several angles varying in severity and originating in different domains. We principally interpreted trustworthiness as being within the bounds of criminal acts (Studies 1–3) or behaviors that covertly deprive or violate others' wants, needs, and opportunities in the interest of self-gain (Study 4): being arrested for breaking a law (Study 1), corporate fraud (Study 2), violent war crimes (Study 3), and cheating to win a cash prize in a laboratory study (Study 4). Although each of these represents one manifestation of behavior that is untrustworthy, together they assess the common idea of

trustworthiness in different ways. If trustworthiness is legible from appearance, then, we would expect one or more of these behaviors to reveal elements of trustworthiness via appearance—similar to the conclusions drawn by earlier work (e.g., Bond et al., 1994; Porter et al., 2008; Stillman et al., 2010; Stirrat & Perrett, 2010; Verplaetse et al., 2007).

Thus, in our first study, we tested the role that context and photo type play in the observation of differences in perceived trustworthiness from faces by comparing judgments of the relative trustworthiness of Nobel Peace Prize winners against that of public figures who had been arrested for breaking the law and had readily available "mug-shot" (arrest) photos and photos from the popular press. In Study 2, we tested trustworthiness in the financial domain by comparing impressions from the faces of corporate criminals versus noncriminal executives. In Study 3, we examined trustworthiness judgments within an aggressive context by testing whether perceived trustworthiness judgments distinguished military servicemen convicted of war crimes linked to illegal prisoner interrogations from controls matched for military rank. In Study 4, we investigated a fairly common form of trustworthiness by exploring whether the faces of students who cheat might be perceived differently in trustworthiness from those who do not cheat and compared these judgments with impressions of other traits known to be accurately perceived (i.e., extraversion and intelligence). Finally, in Study 5, we assessed whether the well-established neural correlates of perceived trustworthiness were related to targets' actual behavior, as well as to perceivers' impressions. The central goal in this work was to examine the previous findings in the literature showing that perceivers show high agreement in their judgments of trustworthiness and then to test the tacit assumption that this consensus is accurate.

## Study 1

To examine whether assessments of trustworthiness are accurate, we first investigated whether stimulus selection could affect trustworthiness judgments. Specifically, it was important to determine whether the type of image (e.g., mug shot versus professional portraits) would sway individuals' judgments. Previous work has examined distinctions between criminals and noncriminals as a proxy for measuring differences in trustworthiness. For instance, Porter et al. (2008) compared perceptions of America's Most Wanted criminals with recipients of major accolades, such as the Nobel Peace Prize and the Order of Canada. Although the authors found that the groups were distinct, the investigation was limited by the incomparability of the two groups; for instance, differences in mug shots versus typical media portraits. In Study 1, we evaluated whether the types of images used would affect trustworthiness judgments. Thus, we also asked participants to judge the trustworthiness of recipients of the Nobel Peace Prize and then compared these judgments with trustworthiness ratings given to photos of relatively obscure minor celebrities who had been arrested (participant recognitions were less than 1%, see below) based on either their mug-shot photographs or their professional media portraits.

### Method

**Stimuli.** Photos of Nobel Peace Prize laureates were downloaded from various Internet sources. We obtained photos of 38 Caucasian male laureates. Each photo was cropped to the extremes of the head (top of hair, bottom of chin, extremes of ears or hair), standardized in size, and converted to grayscale.

Photos of celebrities who had been arrested for various crimes were downloaded from Internet criminal databases (e.g., http://mugshots.com/celebrity). To maintain consistency with the photos of the Nobel Peace Prize laureates, the first 38 photos of Caucasian men returned in the databases were collected. None of the targets encountered were highly famous, as confirmed by the low rate of participant recognitions described below. For each target, a photo of the celebrity was also downloaded from an online media source, chosen by selecting the first photo returned by Google Images in which the individual's face was directly oriented toward the camera, unoccluded, unadorned, and not in costume or in a performance context. The images were standardized using the same procedures as for the photos of the Nobel Peace Prize laureates.

**Procedure.** Eighty-one undergraduates (63% women) rated the photos for trustworthiness along a 7-point scale anchored at 1 (*Not at all trustworthy*) and 7 (*Very trustworthy*) in exchange for partial course credit or monetary compensation. The participants were randomly assigned to one of two conditions: judging the trustworthiness of photos of Nobel Peace Prize laureates and photos of celebrities from media outlets ($n = 41$) or judging the same Nobel Peace Prize laureates and arrest photos of the same celebrities taken from the online criminal databases ($n = 40$). One participant in the latter condition was excluded from analysis for providing uniform responses for all targets at the midpoint of the scale.

Each photo was presented on a computer screen in random order, and participants were not told that the men in the photos were famous or infamous in any way. Most of the targets were only minor celebrities or, in the case of the Nobel Peace Prize laureates, not well known by their appearances. At the end of the experiment, participants were asked to indicate whether they recognized any of the targets and, if so, to type the names of the individuals that they recognized into a text box. Six participants recognized one or more targets in each condition, and so we excluded those trials (0.3% of all data).

**Analysis.** Prior to data collection, we conducted power analyses to assure that we had adequate sample sizes to detect the presence of effects. Specifically, because we were testing for the possibility of null effects (i.e., that trustworthiness judgments may not be accurate), it was critical that we had sufficient statistical power to assure that any absence of a significant effect was not merely an artifact of an underpowered design. The mean effect size (Cohen's $d = 0.61$) was calculated based on the effects reported in Bond et al. (1994; $d = 0.42$), Porter et al. (2008; $d = 0.78$), Stillman et al. (2010; $d = 0.44$), and Verplaetse et al. (2007; $d = 0.80$). Given that we were primarily interested in the comparison of trustworthy versus untrustworthy targets, our main analytic approach was to aggregate ratings across participants to use targets as the unit of analysis. A $t$ test comparing two groups of 38 targets each would yield a power level of only 75% in a two-tailed test; however, the power to replicate the previous effects (i.e., a one-tailed test) with this design would be 84%.

As we were limited in the number of targets that we could obtain, we therefore also analyzed the data using the participants as the unit of analysis. To do this, we calculated sensitivity correlations (e.g., Judd, Ryan, & Park, 1991) by correlating each partic-

ipant's ratings of the targets along the 7-point scale with a dummy-coded vector of 0s and 1s that corresponded to the target group (celebrities and Nobel Peace Prize laureates, respectively). Each participant's sensitivity correlation ($r$) was then converted to a Fisher's $z$ score for analysis. Power calculations indicated that 97% power could be achieved with 41 participants (the sample size of the condition comparing professional photographs of both groups) in a two-tailed test. This means that the probability of falsely accepting the null hypothesis would be $\beta = .03$, less than the standard criterion required for rejecting the null hypothesis ($\alpha = .05$). Conceptually, then, this level of power would allow us to conclude that the null effect is statistically significant at .03. Indeed, our hypothesis was that Nobel Peace Prize laureates would be perceived as significantly more trustworthy than celebrities in their arrest photos, but not in their media photographs.

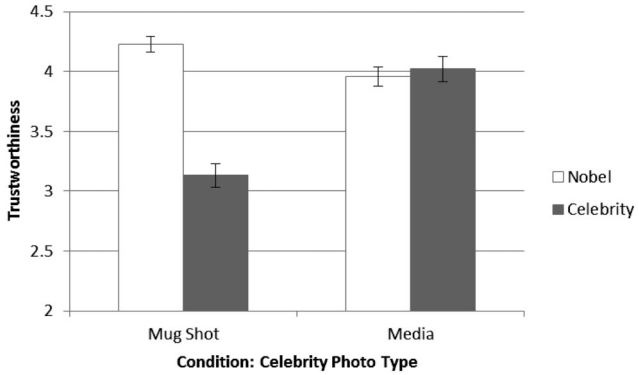


*Figure 1.* Ratings of trustworthiness for the Nobel Peace Prize laureates and celebrities in Study 1 by condition type (celebrity photos were mug shots or photos from the professional media). Error bars represent standard errors of the means.

## Results and Discussion

The participants showed high levels of agreement in their ratings of the targets' faces in both conditions. The mean intercorrelation between the participants' ratings of the targets' trustworthiness in the mug-shot condition was $\bar{r} = .24$ ($SD = .20$) with a 95% confidence interval ranging from .18 to .30, corresponding to Cronbach's $\alpha = .92$; and in the media-photo condition was $\bar{r} = .15$ ($SD = .17$) with a 95% confidence interval ranging from .09 to .20 (Cronbach's $\alpha = .86$). As these confidence intervals do not contain 0, the mean correlations can be regarded as statistically significant at $\alpha = .05$. Thus, the participants showed significant consensus in their impressions of the targets' trustworthiness.

Given the significant agreement between perceivers in their ratings of the targets' trustworthiness, we averaged their judgments across participants to create a mean score for each target. Consistent with the previous research (e.g., Porter et al., 2008), Nobel Peace Prize recipients were rated as significantly more trustworthy ($M = 4.23$, $SE = 0.07$) than were celebrities when photographed as part of their criminal prosecution ($M = 3.13$, $SE = 0.10$): $t(74) = 8.95$, $p < .001$, $d = 2.10$. When the Nobel Peace Prize laureates ($M = 3.96$, $SE = 0.08$) were compared to media photos of the celebrities ($M = 4.02$, $SE = 0.11$), however, the two groups did not significantly differ: $t(74) = 0.48$, $p = .63$, $d = -0.11$; see Figure 1.

Interestingly, these means suggested that the context in which the faces of Nobel Peace Prize laureates were presented (either with celebrity mug shots or celebrity media photos) might have affected how they were judged. To explore this further, we compared the ratings given to the faces across conditions. As expected, celebrity mug shots were rated as significantly less trustworthy than celebrity media photos: $t(37) = 8.06$, $p < .001$, $d = 1.42$. More interesting, however, Nobel Peace Prize laureates were perceived as significantly more trustworthy when evaluated in the context of the celebrities' mug shots than when evaluated in the context of the celebrities' media photographs: $t(37) = 5.42$, $p < .001$, $d = 0.59$. This is consistent with our prediction that the previous effect found for differences between Nobel Peace Prize laureates and America's Most Wanted criminals in Porter et al. (2008) might have been due to the mismatched nature of the photos. Moreover, it illustrates that the same individuals can be given significantly different ratings of trustworthiness depending upon the context created by the other photos rated in the same set (see also Biernat & Manis, 1994). Simply put, trustworthiness ratings may be driven more by the context in which the stimuli are created and assessed, and less by the individual target being judged.

Based on the previous work, the expectation was that Nobel Peace Prize laureates should be rated as significantly more trustworthy than the criminal celebrities in both conditions. A strong test to reject this hypothesis would require a design with at least 95% power, the parallel to $\alpha = .05$, to demonstrate the veracity of a null effect. Given that our items-based analysis was limited by a small number of targets, we therefore conducted a higher power ($1\text{-}\beta = .97$) judge-based analysis using sensitivity correlations. This showed that individual perceivers' accuracy at distinguishing between the Nobel Peace Prize laureates and celebrities' mug shots was significantly above chance (i.e., greater than 0): $M_{\text{Fisher } z} = .40$, $SD = .27$, 95% confidence interval (CI) [.32, .49]. However, individual perceivers' accuracy at distinguishing between the Nobel Peace Prize laureates and celebrities' standard media photographs was not significantly different from chance, as the 95% confidence interval included 0: $M_{\text{Fisher } z} = .01$, $SD = .29$, 95% CI [−.08, .09].

The context in which the photos were perceived therefore appears to have exerted a large effect upon how they were rated in two ways. First, photos of Nobel Peace Prize laureates were rated as significantly more trustworthy than celebrities, but only if the celebrities' photos were from mug shots. Second, the Nobel Peace Prize laureates were rated as significantly more trustworthy when evaluated in the context of mug shots of celebrities. Thus, the context in which a photo is taken influences how trustworthy the target appears. Moreover, the context in which a photo is rated depends on the context created by the other photos in the set, in this case polarizing the ratings that the participants gave. Photo type and evaluation context may therefore have strong effects upon whether individuals are rated as trustworthy or untrustworthy from their faces, perhaps explaining why judgments of trustworthiness from faces were found to be accurate in some past studies (e.g., Porter et al., 2008).

## Study 2

Study 1 showed that photo type and evaluation context can influence individuals' perceived trustworthiness. One conclusion from this is that targets taken from different domains (i.e., Nobel Peace Prize winners versus celebrities arrested for various crimes) may not be compatible matches for comparing relative levels of trustworthiness, as they come from strikingly different fields. In Study 2, we therefore explored whether criminals and noncriminals might be distinguishable when targets were drawn from the same domain: successful business executives.

Financial scandals in business have perhaps become one of the most renowned areas in the popular media in which a lack of trustworthiness is revealed to have major consequences. Moreover, perceivers show high consensus in their perceptions of trustworthiness from executives' faces (Rule & Ambady, 2008). We therefore compared ratings of trustworthiness from the faces of famed corporate criminals, such as Ken Lay of Enron and Dennis Kozlowski of Tyco, against ratings from the faces of their successors or against noncriminal executives from the same industry.

### Method

**Stimuli.** Photos of 15 high-profile business executives who had been convicted of fraud were downloaded from the websites of their companies or from news media outlets. All of the pictures were professional headshot photos of the men in business attire photographed prior to their criminal arrest or indictment, as more recent photos or mug shots may be biased, as indicated by the findings of Study 1; all of the targets were Caucasian men. In addition, we downloaded photos of 15 noncriminal executives of equivalent ranks from the same industries to serve as controls. Control photos were selected by either choosing the criminals' replacements at the same company or individuals of the same rank at a known competitor of the criminal executive's company. The only selection criteria were that the targets were matched in race and sex (i.e., also Caucasian men); otherwise, the photos were selected randomly. Each photo was cropped to the extremes of the head (top of hair, bottom of chin, extremes of ears or hair), standardized in size, and converted to grayscale.

**Procedure.** Fifty undergraduates (Power = 99%) rated the photos on trustworthiness along a 7-point scale anchored at 1 (*Not at all trustworthy*) and 7 (*Very trustworthy*) in exchange for partial course credit or monetary compensation; information about participant sex was lost due to a programming error. Each photo was presented on a computer screen in random order, and participants were not told that the men in the photos were business executives or people who had committed crimes. At the end of the experiment, participants were asked to indicate whether they recognized any of the targets and, if so, to type the names of the individuals that they recognized into a text box; no participants recognized any of the targets.

Because of the influence of affect (Zebrowitz, Kikuchi, & Fellous, 2010) and attractiveness (Dion, Berscheid, & Walster, 1972) on how targets are perceived, we asked independent raters to pretest the faces for differences in emotional expression ($n$ = 20, Cronbach's $\alpha$ = .97) and facial attractiveness ($n$ = 15, Cronbach's $\alpha$ = .89) along 7-point scales anchored at 1 (*Neutral* or *Not at all attractive*) and 7 (*Happy* or *Very attractive*). The two groups of targets did not significantly differ along either variable: $t_{\text{affect}}(28) = 1.60$, $p = .12$, $d = 0.85$; $t_{\text{attractiveness}}(28) = 0.86$, $p = .40$, $d = 0.45$.

### Results and Discussion

The perceivers showed high levels of agreement in their ratings of the targets' faces. The mean intercorrelation between the participants' ratings of the targets' trustworthiness was $\bar{r}$ = .10 ($SD$ = .19) with a 95% confidence interval [.05, .15] that did not contain 0 (Cronbach's $\alpha$ = .86). Perceivers therefore showed consensus in their impressions of the executives' trustworthiness that was statistically significant at $\alpha$ = .05, replicating previous studies showing high consensus in trustworthiness, as discussed above.

Despite this significant consensus, the participants' judgments were not accurate. Aggregating across perceivers, ratings of trustworthiness for the corporate criminals ($M$ = 3.85, $SE$ = 0.10) did not differ from those given to the controls ($M$ = 3.97, $SE$ = 0.14): $t(28) = 0.66$, $p = .51$, $d = 0.24$. Given the small number of prearrest photos of fraudulent executives that we were able to obtain, we sought to conduct a more powerful test for differences between the criminal and noncriminal executives by analyzing the data with the participants as the unit of analysis. As in Study 1, we conducted sensitivity correlations between each participant's ratings and a dummy-coded vector corresponding to group membership. Although the statistical power for this analysis was quite high, the mean sensitivity correlation did not significantly differ from 0: $M_{\text{Fisher's } z} = .04$, $SD = .19$, 95% CI [−.01, .10].

Executives who had been convicted of professional dishonesty therefore were not perceived as significantly less trustworthy than executives not known to be involved in corporate financial scandals. Although the actions of these individuals were dishonest and damaging to the lives of numerous people, perhaps the effects were muted by the white-collar nature of the crimes. It is also possible, though, that the control executives might be as dishonest as the corporate criminals, with the difference between them being less about character and more about who gets caught. To account for some of these limitations, Study 3 investigated a different domain of untrustworthy behavior by comparing judgments made about military criminals convicted of violent war crimes versus military heroes.

## Study 3

Study 2 found that business executives who had been convicted of financial crimes were not perceived significantly differently in trustworthiness compared to business executives not accused of crimes. Although this is a highly impactful domain, the absence of differences between the two groups might have been due to the intellectual or white-collar nature of the untrustworthy act. To broaden our evaluation of trustworthiness judgments and to account for this possible limitation in Study 2, in Study 3 we examined judgments given to targets whose untrustworthy behavior was aggressive and violent. Thus, Study 3 considers perceivers' impressions of the levels of trustworthiness expressed by U.S. military criminals (individuals convicted of committing war crimes) versus U.S. military heroes (individuals receiving the Purple Heart medal).

## Method

**Stimuli.** Photos of 25 U.S. army servicemen recently convicted of war crimes (one or more of murder, rape, conspiracy, or prisoner maltreatment) in military courts and 25 servicemen recently distinguished with the Purple Heart medal were downloaded from U.S. military websites or news media outlets; critically, none of the military criminals had ever been decorated for service. Individuals within the two groups were matched for rank, race, and sex (all male). The photos of the criminals were identified and selected based on news coverage indicating their involvement and conviction of a war crime. The control images were selected by randomly choosing a same-race serviceman of equal rank from military databases of decorated military personnel. Equal numbers of targets from both groups were photographed in military uniforms and civilian clothing; none of the photos were mug shots, and all of the photos were posed photographs. Each photo was cropped to the extremes of the head (top of hair, bottom of chin, extremes of ears or hair), standardized in size, and converted to grayscale.

**Procedure.** Fifty undergraduates (62% women; Power = 99%) rated the photos on trustworthiness along a 7-point scale anchored at 1 (*Not at all trustworthy*) and 7 (*Very trustworthy*) in exchange for partial course credit or monetary compensation. Each photo was presented on a computer screen in random order, and participants were not told that the targets were military servicemen or that they had committed crimes. At the end of the experiment, participants were asked to indicate whether they recognized any of the targets and, if so, to type the names of the individuals that they recognized into a text box; no participants recognized any of the targets. Independent raters coded the faces for emotional expression ($n = 20$, Cronbach's $\alpha = .98$) and facial attractiveness ($n = 30$, Cronbach's $\alpha = .86$) along 7-point scales anchored at 1 (*Neutral* or *Not at all attractive*) and 7 (*Happy* or *Very attractive*). Pretesting showed that the two groups did not differ on either trait: $t_{affect}(48) = 0.02$, $p = .99$, $d = 0.01$; $t_{attractiveness}(48) = 1.18$, $p = .24$, $d = 0.45$.

## Results and Discussion

Although, as in previous work, the judges showed significant consensus in their ratings of trustworthiness ($\bar{r} = .19$, $SD = .17$, 95% CI [.14, .24], Cronbach's $\alpha = .92$), they did not show significant differences in their ratings of the military criminals ($M = 4.01$, $SE = 0.11$) and military heroes ($M = 4.12$, $SE = 0.14$): $t(48) = 0.64$, $p = .53$, $d = 0.18$. Military criminals were therefore not rated significantly differently from military heroes in how trustworthy they appeared from their faces.

Analyses in which participants were the unit of analysis showed similar effects. The mean sensitivity correlation between the individual participants' ratings and a dichotomous vector corresponding to group status (criminals = 0, controls = 1) was within a 95% confidence interval containing 0: $M_{\text{Fisher } z} = .04$, $SD = .13$, 95% CI [.00, .08]. Thus, the judgments were not significantly different from chance.

Studies 2 and 3 therefore suggest that impressions of trustworthiness from faces, despite showing high levels of interrater agreement, may not be diagnostic of targets' actual behavior. Study 4 addressed the question of the legibility of trustworthiness from facial appearance further by photographing students in the lab and then observing the trustworthiness of their behavior.

# Study 4

Studies 2 and 3 asked participants to judge the trustworthiness of business executives and military servicemen who had committed serious crimes. Although these criminal acts certainly characterize behavior that is reprehensible and may therefore be considered untrustworthy, most violations of trust probably do not come from extreme, criminal examples but from discreet acts of dishonest behavior. As untrustworthy behavior leading to criminal incarceration may not represent the norm, Study 4 therefore sought to capture trustworthiness in one of its most common forms: student cheating.

Recent studies have reported that as many as 82% of university undergraduates (McCabe, Treveno, & Butterfield, 2001) and 87% of high school students (Honz, Kiewra, & Yang, 2010) admitted to engaging in cheating behaviors. We therefore created a paradigm that would tempt students to cheat on a test in order to increase their chances of winning a competition for a cash prize. We then asked separate raters to judge the trustworthiness of the cheaters and noncheaters to see whether trustworthiness might be legible from their faces.

## Method

**Stimulus creation.** Forty-six undergraduates ($n = 29$ female) were recruited to participate in a series of experiments in exchange for $10. Upon their arrival in the laboratory, participants were reminded that they would be participating in a series of short experiments and were told that the first study involved the creation of a set of emotional face stimuli to be used in an upcoming study. Participants were asked if they would be willing to be photographed; all agreed. Participants were informed that they would be asked to pose three expressions: happy, angry, and neutral. The participants posing the three expressions were photographed with a digital camera. The photos were taken under conditions standardized for lighting and distance from the camera and with the same neutral, homogeneous background. The angry and happy photos were then discarded, and the neutral photos were retained for later use. Thus, the happy and angry photos served only to support our cover story about the photos being taken for a separate study on emotions and to relax the participants for the final, neutral photo (i.e., pilot testing showed that participants became much less self-conscious about being photographed as neutral after enacting the emotional expressions).

The participants were then told that they would spend the remainder of their time completing a series of computer-based experiments and surveys (i.e., typical psychology experiments). The participants spent approximately 40 min completing these tasks, which consisted of rating faces along various trait dimensions and completing self-report measures of their own traits and personality. The primary purpose of these tasks was to distract the participants and to create a sense of routine normality. Buried within these tasks, however, we asked the participants to respond to two queries of critical interest: (a) to indicate how extraverted they believe themselves to be along a 7-point scale anchored at 1 (*Not at all extraverted*) and 7 (*Very extraverted*) and (b) to respond to the statement "I have never cheated on a test" by rating the statement from 1 (*Not at all true of me*) to 7 (*Very true of me*).

Once they completed all of the computer tasks and surveys, the experimenter informed the participants that there was one, final task in which we needed individuals to pretest math and verbal multiple-choice questions for the development of a test to be used in an experiment the following semester. The participants were told that they would be given a packet of questions from the Graduate Record Examination (GRE) and that we were interested in seeing how many questions people could correctly answer in 5 min. They were told that we would be offering a $100 prize to the person who correctly completed the most questions, in order to assure that everyone tried their hardest and to best simulate the actual upcoming experiment. The experimenter then set a bell-based ticking kitchen timer to the 5-min mark and placed it on the desk next to the participant. The participant was told that he or she should stop working and retrieve the experimenter as soon as the bell rang.

Critically, the participants were being observed through the slats of a set of horizontal blinds that covered a one-way mirror. Participants had performed all of the experiments in the same testing room and, at this point, had been working in the same room for approximately 50 min. From the perspective of an individual inside of the testing room, the mirror was not visible behind the fully closed blinds. However, when the blinds were turned such that the curved/bottom side of the horizontal slats was pivoted 90 degrees downward (such that the slats were vertical), someone on the other side of the glass positioned at a superior height to the participants could see into the testing room to observe their behavior. Thus, from the viewing side of the occluded one-way mirror, the experimenter was able to view the participant working on the questions and was near enough to clearly hear the ticking and ringing of the bell through the glass. The experimenter therefore observed the participant through the mirror and initiated a stopwatch as soon as the bell rang. The experimenter recorded the length of time that the participant worked past the bell to complete the test questions. Approximately half of the participants (57%, $n = 26$) cheated on the test by working beyond the 5-min testing period. Participants were fully debriefed about the nature of all of the tasks in the experimental session and told that the $100 would not go to the person who scored highest on the test but would be administered through a voluntary raffle. All participants chose to enter the raffle and no participants chose to withdraw their photo or data after having been told the full purpose of the experiment, including the experimenter's perception of any cheating behavior.

**Stimulus rating.** Fifty undergraduates (72% female; Power = 99%) at a different university rated the targets' photos in exchange for partial course credit. Participants rated the photos for trustworthiness, intelligence, and extraversion, among other traits, along 7-point scales anchored at 1 (*Not at all X*) and 7 (*Very X*). The traits were rated in randomly ordered blocks within which the presentation of faces was also random. Stimuli were presented, and responses were collected, via computer. At the end of the experiment, participants were asked to indicate whether they recognized any of the targets and, if so, to type the names of the individuals that they recognized into a text box; no participants recognized any of the targets.

## Results

**Consensus.** Judges showed significant agreement in their ratings of the targets from their faces. The mean intercorrelation between the participants' ratings of the targets' trustworthiness ($M = 4.10$, $SE = 0.09$) was $\bar{r} = .20$ ($SD = .18$) with a 95% confidence interval ranging from .15 to .25. This corresponded to Cronbach's $\alpha = .92$. As the confidence interval does not contain 0, the magnitude of this mean correlation can be regarded as statistically significant at $\alpha = .05$. Ratings of intelligence ($M = 3.89$, $SE = 0.11$) yielded a mean interrater correlation of $\bar{r} = .21$ ($SD = .17$), with a 95% confidence interval ranging from .16 to .26, corresponding to Cronbach's $\alpha = .93$. Finally, participants' ratings of extraversion ($M = 4.31$, $SE = 0.09$) were intercorrelated at $\bar{r} = .23$ ($SD = .21$) with a 95% confidence interval of .17 to .28; Cronbach's $\alpha = .94$.

**Accuracy.** Our primary question of interest was whether perceivers' judgments of traits from targets' faces corresponded to the targets' actual behavior. To measure this, we correlated perceivers' trustworthiness ratings with the targets' cheating behavior.[1] We also correlated perceivers' ratings of extraversion and intelligence with targets' own ratings of extraversion and their performance on the GRE test questions, respectively. As above, we conducted these analyses both with the targets as the unit of analysis (by averaging across all participants' judgments) and with the participants as the unit of analysis (using sensitivity correlations); these data are presented together below.

***Accuracy in judging trustworthiness.*** The degree to which perceivers rated the faces as trustworthy along the 7-point scale was statistically unrelated to whether or not the target cheated on the test: $r_{pb}(44) = .06$, $p = .71$; $M_{\text{Fisher } z} = .00$, $SD = .15$, 95% CI [−.04, .04]. In addition, trustworthiness ratings were unrelated to the length of time that targets cheated on the test, wherein noncheaters' time was coded as 0: $r_{\text{Spearman}}(42) = .02$, $p = .90$; $M_{\text{Fisher } z} = .00$, $SD = .12$, 95% CI [−.03, .03]; see Figure 2.[2] The effect was no different when considering the relationship between trustworthiness and cheating time for only the subset of individuals who cheated: $r_{\text{Spearman}}(22) = -.03$, $p = .89$; $M_{\text{Fisher } z} = .01$, $SD = .15$, 95% CI [−.03, .05]. Finally, we separated the targets into bins based on whether they did not cheat (−1, $n = 20$), cheated for less than 30 s (0, $n = 14$), or cheated for more than 30 s (1, $n = 12$) to distinguish the noncheaters, minor cheaters, and more serious cheaters. The goal of this analysis was to present the data in a format compatible with previous research on varying levels of trustworthiness (e.g., Said, Baron, & Todorov, 2009). This variable again showed no relationship to impressions of trustworthiness from the targets' faces, $r_{\text{Spearman}}(44) = .04$, $p = .82$, and the 95% confidence interval surrounding the mean sensitivity correlation contained 0: $M_{\text{Fisher } z} = -.01$, $SD = .15$, 95% CI [−.05, .04].

Interestingly, however, we did observe a relationship between the amount of time that targets cheated and the targets' self-reports of their past cheating behavior (reverse scored): $r_{\text{Spearman}}(42) = -.26$, $p = .09$, albeit marginally significant; see Figure 3. Thus, participants who cheated longer were more likely to deny having

[1] Spearman correlations were used in cases where variables were not normally distributed because the data did not meet the assumption of normality required for the Pearson's correlation.

[2] For two participants, we failed to record the length of time that they cheated but noted that they did work over the time limit. We therefore could not include these targets in the analyses examining cheating time but were able to retain them in the analyses in which targets were coded categorically as cheaters and noncheaters.

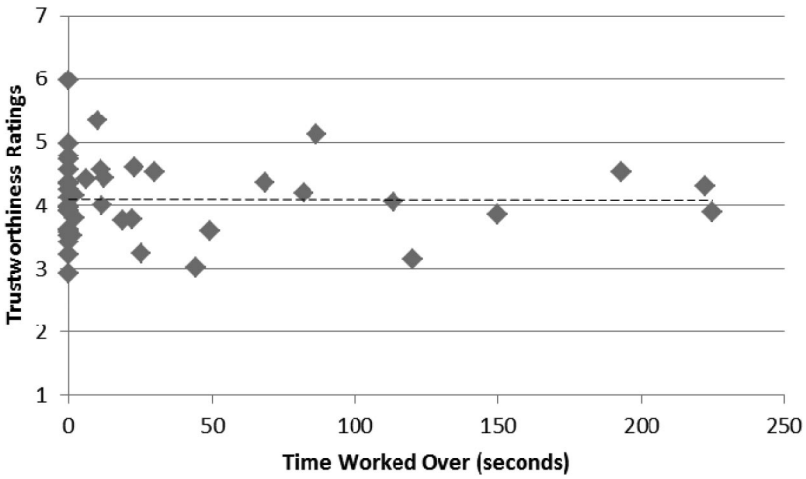


*Figure 2.* Mean trustworthiness ratings of the targets' faces as a function of the duration of time that they cheated (noncheaters' time coded as 0s).

cheated on a test. This finding suggests a connection between a specific behavior (cheating) and a potential trait (trustworthiness) that may be stable across multiple domains. Given that it is only marginally significant, however, this is only tentative evidence that should be explored in future work. Moreover, self-reported cheating was not significantly related to perceived trustworthiness, $r_{\text{Spearman}}$ (44) = .10, $p$ = .52.

***Accuracy in judging extraversion and intelligence.*** Perceivers' impressions of extraversion and intelligence, however, differed from the effects found for trustworthiness. Perceivers' impressions of extraversion from the faces were significantly correlated with the targets' self-reported extraversion, $r_{\text{Spearman}}(43) = .30$, $p$ = .047 (see Figure 4), and the 95% confidence interval surrounding the mean sensitivity correlation did not contain 0: $M_{\text{Fisher } z} = .16$, $SD = .14$, 95% CI [.12, .20].[3] Similarly, perceivers' impressions of intelligence from the targets' faces were marginally correlated with their performance (number of questions answered correctly minus a 20% guessing penalty for each question answered incorrectly) on the test questions, $r_{\text{Spearman}}(44) = .28$, $p$ = .058 (see Figure 5), though the 95% confidence interval surrounding the mean sensitivity correlation did not contain 0: $M_{\text{Fisher } z} = .09$, $SD = .13$, 95% CI [.05, .13]. Previous research has found that the relationship between actual and perceived intelligence was strongly influenced by facial attractiveness (Zebrowitz et al., 2002). When controlling for attractiveness judgments made by a separate sample of participants for the present faces ($N$ = 20; $\bar{r} = .30$, $SD$ − .17, 95% CI [.22, .37], Cronbach's $\alpha = .88$), the relationship between perceived intelligence and performance on the test questions reached significance: $r_{\text{Spearman}}(43) = .33$, $p$ = .03. Based on these findings for extraversion and intelligence, perceivers' failure to accurately judge the targets' trustworthiness in the current study does not appear to be due to the perceivers' inability to extract valid information from the targets' faces but, perhaps, because of an inability to extract valid information about that specific domain of judgment (trustworthiness).

**Manipulation check.** Given that trustworthiness is a broad descriptor of the targets' cheating behavior, we were curious about the relationship between participants' perceptions of trustworthiness and perceptions of cheating behavior, more specifically. We therefore asked a separate sample of 31 undergraduates (61% female; one-tailed power = 95%) to rate each face for the likelihood that the person would cheat on a test from 1 (*Not at all likely*) to 7 (*Very likely*); these ratings were reverse scored for analysis ($M = 4.76$, $SE = 0.10$). The mean intercorrelation between the participants' ratings was $\bar{r} = .20$ ($SD = .16$), 95% CI [.15, .26]), corresponding to Cronbach's $\alpha = .88$. More important, perceptions of trustworthiness and perceptions of likelihood to cheat were significantly correlated: $r_{\text{Spearman}}(44) = .52$, $p < .001$. In addition, when we repeated the above analyses with ratings of perceived cheating in place of ratings of trustworthiness, all of the results remained nonsignificant: all $|r|$s $< .10$, all $p$s $> .52$; all $|M_{\text{Fisher } z}|$s $< .05$, all $SD$s $< .18$, all 95% CIs [−.11, .06].

## Discussion

Despite showing high consensus in their perceptions of individuals' trustworthiness, students who cheated on a test were rated no differently than students who did not cheat. Thus, trustworthiness appears to be no more diagnostic from the neutral faces of everyday people photographed in the lab than it is from the photos of high-profile corporate and military criminals taken from professional and media outlets.

In contrast, extraversion and intelligence did appear to be legible from the students' faces. The perceivers' impressions of extraversion were significantly correlated with targets' self-reports of extraversion. Similarly, perceivers' impressions of intelligence based on neutral photographs corresponded to how well the participants performed on the math and verbal test questions. Interestingly, previous work found that controlling for attractiveness nullified the relationship between perceptions of intelligence and IQ (Zebrowitz et al., 2002), whereas here it seemed to strengthen the effect. This difference could be due to a number of factors (e.g., differences in the target samples, the

[3] One participant's extraversion self-rating was lost due to a computer error.

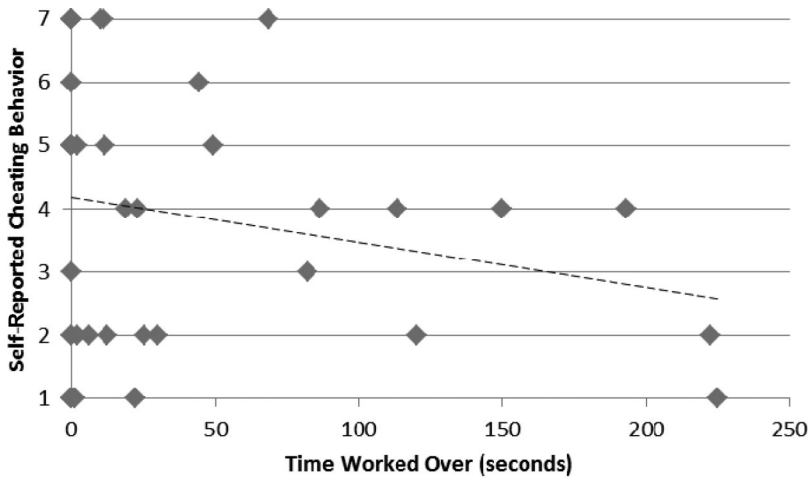


*Figure 3.* Relationship between the amount of time that targets cheated and their self-reported cheating behavior.

use of participants from both Australia and the United States in the previous study, random noise), and additional research would be needed to understand the distinction. Yet, in general, previous work has shown that extraversion (e.g., Penton-Voak et al., 2006) and intelligence (e.g., Zebrowitz et al., 2002) can be reliably judged from the face and the current data support those findings. In turn, previous work has found that perceivers' ability to detect deception from faces is relatively poor (e.g., Bond & DePaulo, 2006). These data would seem to support those findings, as well. In addition, these null effects were further supported by a sample of participants showing that direct judgments of cheating also did not correspond with cheating behavior.

## Meta-Analysis

Thus far, we have shown consistent null effects for judgments of trustworthy versus untrustworthy faces. The means of these effects have been mixed, however. It is therefore possible that, in aggregate, these multiple nonsignificant effects could accumulate to show an overall but modest demonstration of a difference between trustworthy and untrustworthy judgments. To quantify this, we meta-analytically combined the results from the effect sizes reported in Studies 2–4.

In total, we counted 20 effects across the three studies. Of these, nine were negative in sign (suggesting that untrustworthy faces were rated as more trustworthy), eight were positive in sign, and three had no direction (i.e., were equal to 0). Naturally, these 20 effects were not all independent. Rather, they were based on the same data analyzed in different ways (e.g., with faces versus participants as the unit of analysis). We therefore calculated $r$ effect sizes for each of these effects, converted them using the Fisher $z$ transform, and combined the nonindependent effects through averaging. This yielded four independent mean effect sizes, one from each of Studies 2 and 3 and two from Study 4 (one

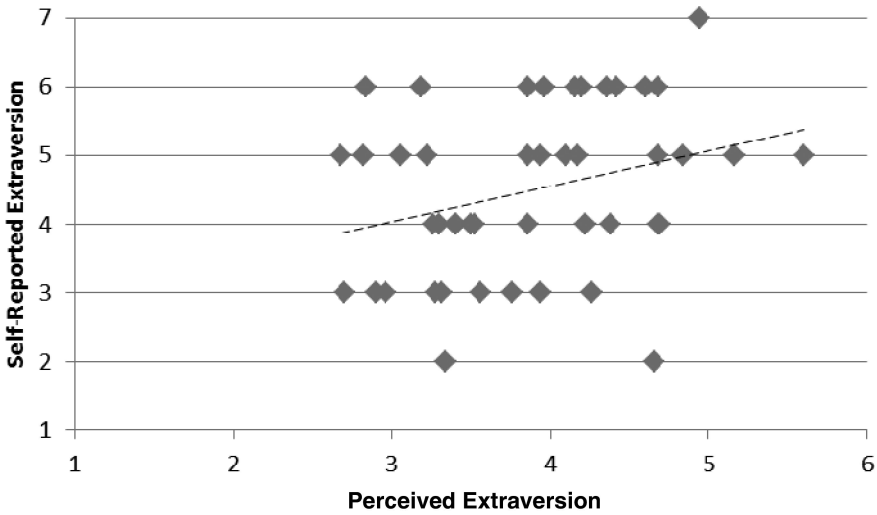


*Figure 4.* Relationship between targets' self-reported extraversion and participants' perceptions of targets' extraversion from photos of their faces.

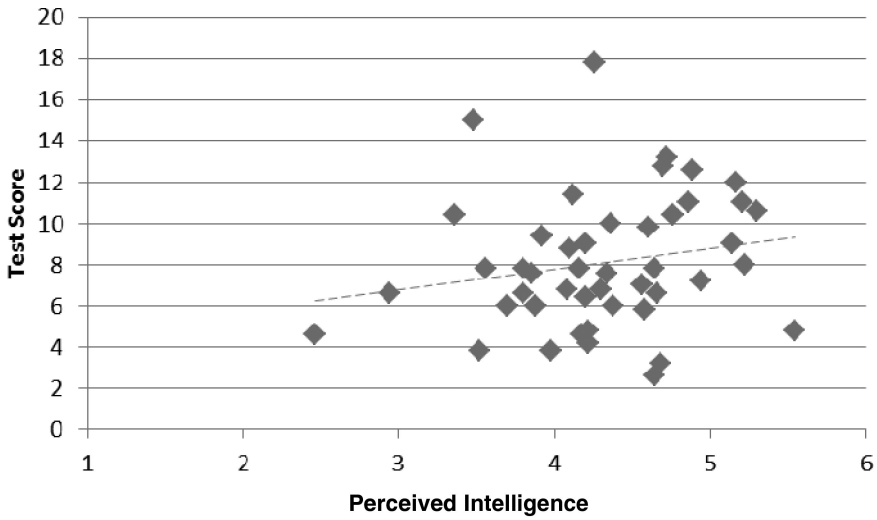


*Figure 5.* Relationship between targets' scores on the test questions and participants' perceptions of intelligence from their faces.

based on the trustworthiness judgments and one based on the cheating judgments in the manipulation check). These effects were then used to obtain a mean overall effect and to construct a 95% confidence interval to test the significance of that effect at $\alpha = .05$. Although the mean effect size was positive ($M_{\text{Fisher's } z} = .03$, $SD = .05$) the confidence interval contained 0, indicating that the overall effect was not significantly greater than chance: 95% CI [−.02, .07].

In sum, subjective impressions of trustworthiness do not seem to correspond to objective measures of trustworthiness. One domain in which impressions of trustworthiness have been extensively studied is in social neuroscience, where the role of the amygdala in response to the trustworthiness of faces has been of great interest (e.g., Adolphs et al., 2002, 1998; Engell, Haxby, & Todorov, 2007; Winston et al., 2002). Therefore, in Study 5, we examined amygdala activation and judgments of trustworthiness.

## Study 5

An emerging body of research examining the neural correlates underlying implicit social cognition has implicated the amygdala as fundamental to social perception. Indeed, the amygdala has been shown to respond to a wide variety of implicit social cues, including fearful faces (Whalen et al., 1998), variations in eye gaze (Adams, Gordon, Baird, Ambady, & Kleck, 2003), and even indications of success (Rule et al., 2011). The amygdala also plays a central role in evaluating the relative trustworthiness of a face (e.g., Adolphs et al., 2002, 1998; Engell et al., 2007; Winston et al., 2002). Importantly, the amygdala is active in response to high-consensus trustworthy and high-consensus untrustworthy faces when perceivers make either implicit (e.g., Todorov, Baron, et al., 2008) or explicit (e.g., Said et al., 2009) judgments of the face. Thus, the amygdala is critical to social perception, especially in determining whether or not someone appears to be trustworthy.

Interestingly, consensus ratings of trustworthiness have been shown to better predict amygdala activation than individual judgments (Engell et al., 2007). Further, the amygdala has been found to respond in a nonlinear fashion to faces perceived as trustworthy (among other traits; Liang, Zebrowitz, & Zhang, 2010; Said, Dotsch, & Todorov, 2010; Winston, O'Doherty, Kilner, Perrett, & Dolan, 2007), such that high-consensus trustworthy faces and high-consensus untrustworthy faces elicit the greatest amygdala activation (Said et al., 2009), whereas faces perceived to be moderate in trustworthiness elicit less activation.

The goal of Study 5 was to examine amygdala responses to faces of individuals who had shown untrustworthy and trustworthy behavior in Study 4. To this end, we focused on two central analyses: first, we sought to replicate previous findings demonstrating that the amygdala responds more to faces perceived as extreme in trustworthiness compared to faces perceived as moderate in trustworthiness; second, we investigated whether the amygdala may be more responsive to the faces of people according not to how trustworthy they are perceived to be but according to how trustworthy they had actually behaved in the lab.

Although the findings of Study 4 showed no relationship between explicit judgments of trustworthiness and targets' actual trustworthy behavior, given that the amygdala has been found to respond to implicit cues to trustworthiness (e.g., Winston et al., 2002), it remained possible that unconsciously detected facial characteristics might still stimulate an amygdala response. Indeed, implicit evaluations of other people have been shown to influence participants' judgments without their awareness (e.g., Macrae, Milne, & Bodenhausen, 1995). Considering the large neuroimaging literature on trustworthiness, we thought it important to investigate whether these judgments are accurate as a means to clarify what the established, implicit amygdala response to face trustworthiness truly represents.

### Method

A total of 18 neurologically normal right-handed female undergraduates were recruited from the greater Boston area to

participate in exchange for monetary compensation. One participant was excluded due to excessive movement during the scanning session (>2 mm), and three additional participants were excluded because they had recognized more than 20% of the target faces, leaving 14 remaining participants. Anatomical and functional whole-brain imaging was performed on a 3.0-T Siemens Tim Trio Scanner using standard data acquisition protocols. Anatomical images were acquired using a high-resolution 3-D magnetization-prepared rapid gradient echo sequence (MP-RAGE; 144 sagittal slices, TE = 7 ms, TR = 2,200 ms, flip angle = 7°, 1 × 1 × 0.89 mm voxels). Functional images were collected in one functional run of 100 time points, using a fast field echo-planar sequence sensitive to blood-oxygen level-dependent contrast (T2*; 31 axial slices per whole-brain volume, 3-mm in-plane resolution, 4-mm thickness, 0-mm skip, TR = 2,000 ms).

**Behavioral task.** While in the scanner, participants viewed the 46 faces from Study 4 for 2 s each. Each face was presented once, in an order that was counterbalanced across participants. Rather than present the faces multiple times, which would increase power, we chose to present each face only once to avoid habituation to the faces that might cause a loss in amygdala sensitivity (e.g., Hart et al., 2000). Participants were instructed to indicate via button press whether they thought the face was symmetrical or asymmetrical. The goal of this task was to ensure that participants attended to the faces without making explicit judgments of trustworthiness (see Rule et al., 2011; Winston et al., 2002). The faces were presented in pseudorandom order in an event-related fashion with jittered fixation throughout.

As the stimuli from Study 4 were created using targets from the same geographic area as the participants in Study 5, we asked participants after the scan to view each of the faces on a computer and to indicate via button press whether they recognized the target from outside of the experiment. We then created separate regressors for each participant that modeled the trials in which a recognized face was seen, and these regressors were used as nuisance variables so as to exclude the recognized faces from analysis.

**Imaging data analysis.** The fMRI data were analyzed using the general linear model for event-related designs in SPM8 (Wellcome Department of Cognitive Neurology, London, England). Data underwent standard preprocessing to remove sources of noise and artifact. Functional data were spatially smoothed (8-mm full-width-at-half-maximum [FWHM]) using a Gaussian kernel.

Our main question of interest in this study was to determine whether participants showed a unique pattern of neural activity in response to making implicit judgments of the faces of cheaters versus noncheaters. In other words, do participants show any neural evidence (specifically in the amygdala) of detecting whether someone will engage in an untrustworthy (cheating) or trustworthy (not cheating) act? Before examining that question, however, we first needed to verify that our data replicated the numerous previous findings in the literature suggesting that the amygdala is more active to faces that are perceived to be highly trustworthy and highly untrustworthy.

To determine whether the amygdala was more active to high-consensus trustworthy and high-consensus untrustworthy faces, we transformed the consensus trustworthiness ratings for each face (obtained in Study 4) from linear to quadratic. We then used both the linear and quadratic ratings as simultaneous parametric regressors. The linear and quadratic ratings therefore served as continuous orthogonal regressors onto which we could regress brain activity. Specifically, if amygdala activity is driven primarily by perceptions of high or low trustworthiness (instead of just low trustworthiness; e.g., Winston et al., 2002), then the amygdala would show heightened activation in response to a quadratic regressor of trustworthiness even when controlling for the separate linear effects of trustworthiness.

Using orthogonal regressors is the most conservative approach to identifying amygdala activity. We therefore also chose to use a more liberal approach based on that described by Said et al. (2009). In this analysis, we used the consensus trustworthiness ratings obtained in Study 4 to separate our faces into four bins corresponding to highly trustworthy, moderately trustworthy, moderately untrustworthy, and highly untrustworthy. We created the bins by taking the mean trustworthiness ratings from Study 4 and converting them to *Z* scores. We then divided the faces into four bins of 11 faces each (i.e., the faces that were more than 1 standard deviation above and below the mean were placed in two different bins as the extreme high and low trustworthy faces, respectively, and the faces that were within 1 standard deviation of the mean were placed in two separate bins as the moderate trustworthy and moderate untrustworthy faces). We excluded the two faces surrounding 0 in order to have an equal number of faces in each bin. If amygdala activity emerged in both the most conservative and relatively more liberal analyses, we could be confident of our results.

As mentioned above, we created a separate regressor for the trials in which participants saw faces that they recognized and excluded these faces from the subsequent analyses. We also created separate regressors for covariates of no interest (a session mean, a linear trend, and six movement parameters derived from realignment corrections) to compute parameter estimates (β) and *t*-contrast images (containing weighted parameter estimates) for each comparison at each voxel for each subject.

Our second interest in these analyses was to determine whether the amygdala distinguished between actual trustworthy and untrustworthy behavior (e.g., people who cheated or did not cheat in Study 4, respectively). To this end, we compared the neural activation to the faces of the cheaters against activation to the faces of the noncheaters. As before, we created a separate regressor for each participant for faces that were recognized during the task and excluded these faces from the final analysis. To determine if amygdala activity distinguished between highly untrustworthy and moderately untrustworthy behavior, we binned the people who cheated into two groups: relatively high and relatively low cheaters.

In Study 4, we evaluated whether people cheated by working beyond the set time limit. For those who did work beyond the time limit, we measured how long beyond the limit they worked. Thus, this second measure gave us a quantifiable measure of how much people cheated. We used these data to bin our targets by those who were "high cheaters" versus those who were "low cheaters." It should be noted that, due to the nature of the task, the reverse contrast (people who did not cheat a little versus people who did not cheat a lot) was not possible. As mentioned above, we did not

have a continuous measure of cheating time for two targets, so they were not included in this analysis. The remaining 24 faces equally divided into two bins of low cheaters and high cheaters, as in Study 4. For all contrasts, average parameter estimates were extracted using anatomically defined masks of the left and right amygdala for the task > baseline contrasts.

## Results

**Amygdala activity as a function of regressing linear and quadratic trustworthiness ratings for individual targets.** To determine whether the amygdala was more active to high-consensus trustworthy and high-consensus untrustworthy faces, we first used a quadratic regressor of perceived trustworthiness ratings to identify amygdala activity. Specifically, if amygdala activity is driven primarily by perceptions of high and low trustworthiness (instead of just perceptions of low trustworthiness), then the amygdala should show heightened activation in response to a quadratic regressor of perceived trustworthiness even when controlling for the separate linear effects of trustworthiness.

Since we had an *a priori* interest in the amygdala, we examined the results of the parametric analysis at $p$-uncorrected $< .01$ with a 5-voxel extent-threshold ($k$) using a small volume correction for the (anatomically defined) bilateral amygdala. Our main interest in this analysis was whether we would see heightened amygdala activity in response to the quadratic regressor when controlling for the linear regressor. Indeed, results revealed activity in bilateral amygdala [left: $-24, 0, -12$; $t(13) = 2.81$, $p < .01$, $r = .62$; right: $27, -3, -18$; $t(13) = 3.20$, $p < .01$, $r = .66$; see Table 1 for complete list of activations[4]], whereas no activity was present at this threshold for the linear regressor when controlling for the quadratic regressor; bilateral amygdala: $t(13) = 0.69$, $p$-uncorrected $= .25$, $r = .19$. Thus, the results from this analysis suggest that the amygdala is uniquely active to high-consensus trustworthy and high-consensus untrustworthy faces.

**Amygdala activity as measured by binning faces according to perceived trustworthiness.** We then conducted a modified version of the paradigm used by Said et al. (2009) in which we divided the faces into four groups corresponding to highly untrustworthy, moderately untrustworthy, moderately trustworthy, and highly trustworthy, based on perceivers' consensus. Although this analysis was relatively more liberal than the regressor analysis described above, we used this alternate approach to provide converging evidence that the amygdala is involved in detecting faces that are perceived to be highly trustworthy and highly untrustworthy. Since we had an *a priori* interest in the amygdala, we examined the group analyses for each bin at $p$-uncorrected $< .01$ with a 5-voxel extent-threshold ($k$). Consistent with Said et al. (2009), we observed the most robust responses in the bilateral amygdala for those faces falling into the highly untrustworthy and highly trustworthy groups (see Figure 6).

To more closely examine the extent of amygdala activation for each of the four groups, we conducted a region of interest (ROI) analysis on the left and right amygdala (anatomically defined) to extract the mean signal change for each of the four groups in their respective face > baseline conditions. We then compared these values using a repeated-measures analysis of variance (ANOVA) for participants' responses to the faces in the four groups for each of the left and right amygdala. Results for the left amygdala showed a significant omnibus effect between the four groups, $F(3, 39) = 3.24$, $p = .03$, $\eta^2_{\text{partial}} = .20$, which was further characterized by a significant quadratic trend: $F(1, 13) = 8.18$, $p = .01$, $r = .62$. The omnibus effect for the right amygdala was only marginally significant, $F(3, 39) = 2.72$, $p = .06$, $\eta^2_{\text{partial}} = .17$, but also showed a significant quadratic contrast: $F(1, 13) = 5.91$, $p = .03$, $r = .56$. These findings are similar to what has been reported in previous research (Said et al., 2009). Consistent with those authors' interpretation, faces perceived as high and low in trustworthiness elicited a greater response in the bilateral amygdala than did the faces moderate in trustworthiness.[5]

**Amygdala activity as measured by actual trustworthiness.** Next, we examined whether the amygdala responded to people's actual behavior. As in Study 4, we first compared whether the amygdala responded differently to the faces of people who cheated versus those who did not, which showed no difference at a threshold of $p$-uncorrected $< .01$, $k = 5$ voxels, which is the same threshold used for the analyses of perceived trustworthiness. We then chose to parallel the binned analysis that we had conducted with the trustworthiness ratings to determine if amygdala activity was present even in this more liberal analysis. Thus, we divided the data into three groups: high cheaters, low cheaters, and noncheaters. We modeled participants' responses to these faces as compared to baseline; none showed a significant response in the amygdala at $p$-uncorrected $< .01$, $k = 5$ voxels. Simply put, the amygdala was not responsive to differences in the trustworthiness of individuals' behavior, only to the perception of individuals' trustworthiness.[6]

## Discussion

Here we found that the amygdala responded to faces that were perceived to be more or less trustworthy but did not differentiate between the faces of people who had engaged in trustworthy and untrustworthy behavior. Thus, the amygdala responded according to the perceived trustworthiness of faces but not to the actual trustworthiness of individuals' behavior.

Generally speaking, one would expect that perception would be a necessary intermediary between an observation and a neural response. Thus, as we did not observe a relationship between individuals' actions and how trustworthy they were perceived to be

[4] It is important to note that the results listed in Table 1 reflect the regions that are active at a threshold of $p$-uncorrected $< .01$ and therefore should be interpreted with caution. This relatively liberal level was selected in order to isolate activation in the amygdala (which was subsequently small-volume corrected) and, as a result, may have presented misleadingly high $t$ values for the remaining areas.

[5] Results were similar when excluding the faces of the two participants for whom we failed to record the duration of cheating time. The left amygdala showed a significant omnibus effect between the four groups, $F(3, 39) = 3.63$, $p < .03$, $\eta^2_{\text{partial}} = .22$, which was further characterized by a significant quadratic effect: $F(1, 13) = 7.85$, $p < .02$, $r = 61$. The omnibus effect for the right amygdala was only marginally significant, $F(3, 39) = 2.48$, $p = .08$, $\eta^2_{\text{partial}} = .19$, but also showed a significant quadratic contrast: $F(1, 13) = 5.71$, $p = .03$, $r = .55$.

[6] To determine whether any of the other neural regions listed in Table 1 predicted cheating, we conducted a region of interest analysis comparing activations to people who cheated versus those who did not; results showed no differences.

Table 1
*Brain Areas Responding to the Quadratic Regressor of Perceived Trustworthiness in Study 5*

| Brain region | Talairach coordinates | | | *k* extent | *t* score |
|---|---|---|---|---|---|
| | x | y | z | | |
| Left amygdala[a] | −24 | 0 | −12 | 12 | 2.81 |
| Right amygdala[a] | 27 | −3 | −18 | 48 | 3.20 |
| Right middle temporal gyrus (BA 21) | 51 | −42 | −3 | 721 | 8.69 |
| Left cerebellum | −21 | −51 | −30 | 932 | 6.18 |
| Left inferior parietal cortex (BA 40) | −57 | −54 | 30 | 427 | 6.02 |
| Right postcentral gyrus (BA 2) | 60 | −18 | 42 | 75 | 5.88 |
| Right precentral gyrus (BA 4) | 27 | −21 | 45 | 40 | 5.73 |
| Right uncus (BA 20) | 30 | −18 | −39 | 428 | 5.37 |
| Right inferior frontal gyrus (BA 47) | 48 | 21 | 3 | 46 | 5.36 |
| Left cerebellum | −21 | −27 | −27 | 127 | 5.10 |
| Right superior frontal gyrus (BA 10) | 27 | 63 | 18 | 293 | 5.02 |
| Paracentral lobule (BA 4) | 0 | −45 | 72 | 164 | 4.71 |
| Left middle temporal gyrus (BA 38) | −33 | 9 | −42 | 108 | 4.67 |
| Left cuneus (BA 18) | −6 | −78 | 15 | 429 | 4.16 |
| Left cingulate gyrus (BA 24) | −6 | −18 | 42 | 36 | 4.10 |
| Left transverse temporal lobe (BA 41) | −36 | −33 | 9 | 51 | 4.08 |
| Right inferior frontal gyrus (BA 47) | 30 | 24 | −15 | 47 | 3.98 |
| Left precentral gyrus (BA 6) | −60 | −15 | 42 | 131 | 3.96 |
| Left inferior frontal gyrus (BA 47) | −42 | 24 | 3 | 127 | 3.91 |
| Right precentral gyrus (BA 6) | 48 | −6 | 39 | 49 | 3.87 |
| Right midbrain | 3 | −12 | −21 | 57 | 3.79 |
| Left superior temporal gyrus (BA 38) | −24 | 15 | −30 | 126 | 3.77 |
| Left precuneus (BA 19) | −33 | −69 | 39 | 59 | 3.75 |
| Left midbrain | −6 | −9 | −6 | 49 | 3.66 |
| Left precentral gyrus (BA 9) | −45 | 18 | 42 | 65 | 3.43 |
| Right superior temporal gyrus (BA 22) | 51 | −60 | 12 | 22 | 3.36 |
| Left middle temporal gyrus (BA 20) | −48 | −36 | −15 | 15 | 3.02 |

*Note.* All imaging data reported were calculated at $p$-uncorrected $< .01$. We used a Monte Carlo conversion script from Slotnick, Moo, Segal, & Hart (2003) to determine the extent threshold required to convert $p$-uncorrected $< .01$ to $p$-corrected $< .05$. We chose a 1,000-iteration Monte Carlo resampling to select the most conservative threshold (13-voxel extent-threshold). The corrected results ($p$-corrected $< .05$, 15-voxel extent threshold) are reported here. BA = Broadmann Area.
[a] Amygdala activity shown at $p$-uncorrected $< .01$, with a small volume correction.

in Study 4, it would seem reasonable that we would not observe an amygdala response as a function of the targets' cheating behavior. Yet, a growing body of evidence has shown that the brain often responds in the absence of conscious perception. For instance, blind-sight patients show response patterns in the visual and temporal cortices, including the amygdala, without awareness of seeing anything (Vuilleumier et al., 2002); subliminal presentations of fear-inducing stimuli elicit a response in the amygdala (Whalen et al., 1998), and retinal-collicular-pulvinar pathways from the optic nerve can provide input to structures such as the amygdala before perceptual information is processed by the visual cortex and other systems (Morris, Ohman, & Dolan, 1998; Vuilleumier, Armony, Driver, & Dolan, 2003). Moreover, there is also evidence that higher order cortical structures, such as regions related to executive function, can suppress or inhibit perceptions in the amygdala—suggesting that it could be possible for the amygdala to respond to individuals' actual trustworthiness but that such a response could be dampened or augmented by the outputs of executive processing (e.g., Bishop, Duncan, Brett, & Lawrence, 2004). Thus, it would not have been unreasonable to expect an amygdala response to individuals' behavior despite a disconnect between that behavior and perceivers' conscious perceptions of the targets' trustworthiness. We did not find such a relationship between targets' behavior and perceivers' amygdala response, however. Rather, perceivers' impressions of trustworthiness alone corresponded to activity in the amygdala.

## General Discussion

People show strong agreement in their perceptions of trustworthiness. Despite this high consensus, however, these perceptions do not seem to correspond strongly with how individuals actually behave. We found that perceivers agreed in the levels of trustworthiness that they ascribed to individuals representing a broad swath of untrustworthy behaviors: various crimes (Study 1), corporate fraud (Study 2), war crimes (Study 3), and cheating on a test in the lab (Study 4). In none of these cases did we find that impressions of who was trustworthy and who was untrustworthy correlated with actual behavior. Finally, we found that although individuals' perceptions of trustworthiness were reliably associated with a response in the amygdala when perceiving targets' faces, this signal did not relate to the trustworthiness of the targets' behavior (Study 5).

Previous research has reported that humans tend not to perform well at detecting deception (Bond & DePaulo, 2006, 2008). Although much of the work contributing to this literature has focused on perceivers' abilities to ascertain others' honesty from specific behaviors (e.g., eye-contact; DePaulo et al., 2003; Sporer & Schwandt, 2007), the current work measured more general assess-

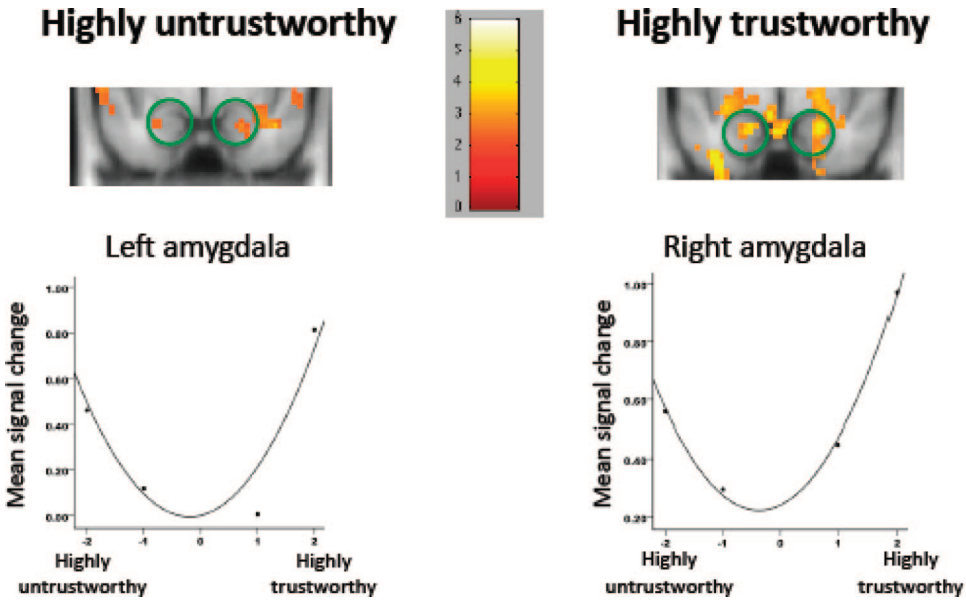


*Figure 6.* Statistical parametric maps and line graphs of left and right amygdala showing a significant quadratic effect [i.e., greater responses to faces perceived to be highly (un)trustworthy compared to faces perceived to be moderately (un)trustworthy] during the perception of faces varying in perceived trustworthiness in Study 5. The color-coded bar indicates *t* values for the contrast analyses. Imaging results reported at *p*-uncorrected $< .01$, $k = 5$ voxels.

ments of targets' overall character. Interestingly, in Study 4 we found that actual cheating in the lab was inversely related to self-reports of typical cheating behavior. If we were to accept this at face value, it would suggest that more honest participants in Study 4 were inclined to cheat to a greater extent in the experiment. Alternatively, it may reinforce the more likely interpretation that the more an individual cheated, the more likely he or she was to lie about cheating.

Perceivers were not inaccurate in their assessments of all traits, however. In contrast to trustworthiness, perceivers' impressions of the targets' extraversion and intelligence in Study 4 were verified by reliable external criteria. Specifically, individuals who self-reported as more extraverted were perceived by others to indeed be more extraverted. In addition, individuals who performed better on the test questions in Study 4 were perceived by others to be more intelligent than those who performed worse. Notably, these targets were the same individuals who were misperceived with regard to their trustworthiness by the same perceivers. This inconsistency suggests that the domain of judgment may account for these differences in judgmental accuracy. Thus, it is not that the targets were illegible or that the perceivers were poor judges but, perhaps, that trustworthiness is not easily judged. One aspect that may contribute to this is differences in the nature of the traits: intelligence may index an ability, extraversion relies principally on consensus or self-report, and trustworthiness seems to be most commonly defined by behaviors.

Previous theoretical models might help to account for this difference. One such model, Funder's (1995) Realistic Accuracy Model, parses the relationship between perception and accuracy into four stages: relevance (the trait must be manifested in some relevant behavior), availability (that behavior must be observable), detection (the perceiver has to be able to detect the behavior), and utilization (the perceiver must link the detected behavior back to the initial trait). A failure at any point in this process can be sufficient to prevent an accurate judgment. Thus, it would seem that extraversion and intelligence meet each of these requirements but trustworthiness does not. One can speculate about why trustworthiness does not satisfy these stages; for example, perhaps information regarding trustworthiness is not as strongly associated with behavioral or self-reported cues as are other traits (e.g., intelligence or extraversion), is only relevant in the moment of an untrustworthy act (as in Verplaetse et al., 2007), fluctuates in availability depending on a person's sex and age (as perhaps seen in Zebrowitz et al., 1996), is detected only by judges with particular skills or training (e.g., O'Sullivan, 2008), or is only utilized by perceivers when it is salient to their own success or failure (as in economic games; see Stirrat & Perrett, 2010). A greater consideration of the body of work on trustworthiness would be needed to elucidate which stages of the Realistic Accuracy Model are satisfied in which contexts. Doing so would help to parse whether the absence of an association between facial appearance and trustworthy behavior in the present work was the consequence of a failure of perceivers to accurately detect cues to trait trustworthiness, a failure of targets to validly express cues to their trustworthiness, or a combination of both (see also Brunswik, 1956).

A second theoretical perspective relevant to the current effects is the Ecological Model of Social Perception (McArthur & Baron, 1983; Zebrowitz & Collins, 1997). An offspring of Gibson's Ecological Theory of object perception, the ecological approach to social perception places the functional utility of percepts as central to understanding the perceptual process. In simple reduction, we attend to and are able to readily observe

aspects of social behavior that are useful to us. Although many have convincingly argued that there is high adaptive value in being able to ascertain who is trustworthy and who is not (e.g., Cosmides, 1989), a potential countervailing force is that there is also high adaptive value in being able to disguise one's true feelings and intentions. Thus, there is value for perceivers in assessing trustworthiness but also value for targets to disguise their (un)trustworthiness. This may be less applicable for traits like extraversion and intelligence, perhaps explaining why they were correctly perceived and trustworthiness was not. Accurately expressing extraversion can be useful for social functioning but is valued differently across cultures (Searle & Ward, 1990). Thus, while it is useful to know whether others are introverted or extraverted, there is not necessarily universal, evolutionary value in trying to pass as extraverted or introverted, as is thought to be the case for trustworthiness (Cosmides, 1989). Similarly, one could argue for adaptive utility in being able to accurately assess others' intelligence, but there is little pressure for the intelligent to pass as unintelligent, and successfully passing as intelligent would seem to require enough cognitive capacity and insight to question whether such individuals are truly unintelligent at all (e.g., McClelland, 1973). Thus, there are differences in these traits at many levels, and although the present data cannot offer a clear and definite explanation for why trustworthiness is not legible from facial cues, it can help to revise our understanding of why some previous studies have reported that it is legible.

Our findings differ from some work examining the judgment of trustworthiness in the context of economic games. Verplaetse et al. (2007) and Stirrat and Perrett (2010) both found that judgments of traits from individuals' faces were related to their behaviors during economic games. Verplaetse et al. (2007) found that perceivers could reliably judge targets' cooperativeness from photos taken during the decisionmaking moment of a prisoner's dilemma game (but not for photos taken before the game). Similarly, Stirrat and Perrett (2010) found that participants' behavior in an economic game was significantly related to their facial width:height ratio (individuals who were more likely to defect against their imagined partner in the game had wider faces) and showed that manipulations of width:height ratio influenced perceivers' impressions of trustworthiness. We believe that this difference may be due to the way that trustworthiness has been operationalized. In Stirrat and Perrett's (2010) and Verplaetse et al.'s (2007) studies, the measure of trustworthiness (cooperativeness in an economic game) could be alternatively interpreted as a measure of agreeableness, lack of selfishness, lack of aggression, or cooperativeness. In fact, similar work has reported just those effects: Carré and McCormick (2008) found that facial width:height ratio was a reliable indicator of individuals' aggression. Thus, given that aggression and trustworthiness are very highly correlated (e.g., $r = -.90$ in Carré et al., 2009), it may be that a lack of aggressiveness can serve as a proxy for trustworthiness in actual behavioral interactions under some circumstances. Further work will clearly be needed to fully disentangle the relationship between trustworthiness, aggression, and behavior.

Even if trustworthiness judgments from faces do not accurately reflect the way that people behave, they still provide useful social information. Indeed, it is an interesting question to consider why perceivers show high agreement in their perceptions of who is trustworthy absent a correspondence with observed trustworthy and untrustworthy behaviors. Previous theoretical models may help to explain why some people are perceived as (un)trustworthy in spite of their actual behavior. Overgeneralization theories, for example, suggest that inferences of trustworthiness may be linked to perceptions of emotional expressions (Zebrowitz et al., 2010; Zebrowitz & Montepare, 2008). Specifically, faces resembling positive emotions (such as happiness) may be seen as approachable and trustworthy, whereas faces resembling negative emotions (such as anger) may be seen as unapproachable and untrustworthy (Oosterhof & Todorov, 2008; Todorov, Said, et al., 2008). Thus, even though judgments of trustworthiness appear not to be accurate in relation to individuals' observed behavior, they can still be important because they represent people's reputation or how they are seen by others.

One limitation of both the past and present work is the undefined nature of what it means to be "trustworthy" or "untrustworthy." Previous researchers have drawn conclusions about trustworthiness using various methods and manipulations. This has resulted in some mixed effects, with some scholars finding differences suggesting that trustworthy and untrustworthy individuals can be distinguished (e.g., Bond et al., 1994), while others have not (Zebrowitz et al., 1996). Moreover, the neuroimaging studies that have examined perceived trustworthiness have operationalized trustworthiness in terms of perceiver agreement. To address this issue, we attempted to assess a broad array of behaviors in the current work. Although we consistently found null effects across each of these domains, it would be of benefit to the field to hone an understanding of "trustworthiness" and its various subtypes. Such an effort should consider unpublished manipulations of trustworthiness. Only with a complete picture can we resolve the heterogeneity in findings thus observed in the literature.

A related limitation concerns whether trustworthiness can be considered a trait- or state-level variable. The work by Zebrowitz et al. (1996) examined the perception and accuracy of trustworthiness across development. They found consistency in some cases but not in others; for example, women were found to show an "artifice effect" wherein lower levels of actual trustworthiness predicted higher levels of perceived trustworthiness later in life (i.e., they learned how to fool people over time), whereas men showed some suggestion of a "self-fulfilling prophecy" whereby honest men grew to have honest faces. Verplaetse et al. (2007) also reported inconsistencies: participants "revealed" their untrustworthy nature only at specific, key time-points during their playing of an economic game but not when photographed before the game or at other points within the game. Yet the findings of Stillman et al. (2010); Porter et al. (2008), and Bond et al. (1994) hint that some cues to trustworthiness should be stable and trait-based, an idea supported by studies suggesting that individuals who cheat are more likely to cheat again (Davis & Ludvigson, 1995; see also Rotter, 1980). If so, it raises an interesting question as to why a static face may contain valid cues to trustworthiness to begin with, a topic that we do not speculate about here given that our results do not promote that conclusion (but see Zebrowitz et al., 1996, for a thorough discussion). This is also relevant to considerations of both overgeneralization theories and the potential physical manifestation of trustworthiness in the face. The facial

width:height ratio differences found in previous work (Stirrat & Perrett, 2010) would suggest a permanence about trustworthiness in the face (see also Malatesta, Fiore, & Messina, 1987). Yet the theoretical models suggesting that trustworthiness may be an overgeneralization of emotional displays (e.g., Oosterhof & Todorov, 2008) also highlight the potential for variance in perceived trustworthiness due to emotional expression (though this is not the message of the overgeneralization theories, which would actually do more to support a permanence interpretation). Thus, clarity around these differences is needed to ascertain a more functional understanding of trustworthiness and of when it is and is not expected to be accurate.

Related to this, it is notable that the majority of behaviors assessed in the current work were criminal acts. Although we sought to converge upon an underlying latent construct of trait trustworthiness by examining behaviors that varied widely in the domain and severity of untrustworthiness, the present work is limited in that it necessarily focused on a finite number of behaviors. Additionally, we and past researchers have assumed that single behaviors may provide credible measures for general trustworthiness, whereas it is possible that trustworthiness is fairly domain specific. For example, people who embezzle millions of dollars from their companies could ironically be very honest about paying their bus fare. A common assumption made here and elsewhere is that trustworthiness in one domain is likely to be correlated with trustworthiness in other domains, such that people who do not cheat on their taxes would also not be expected to cheat on their spouse. Although we did find some evidence of consistency across types of cheating in Study 4 (target-participants who cheated were also more likely to lie about having cheated; or perhaps deceived themselves into believing that they had not cheated, which might also be considered a marker of untrustworthiness, see von Hippel & Trivers, 2011), this remains an empirical question yet to be measured. A better understanding of the consistency of trustworthiness within a single individual across domains might help to alleviate some of the confusion in the trustworthiness and appearance literature that may stem from the use of state-level variables to infer trait-level behaviors. Doing so could help to illuminate some of the inconsistencies in the data on trustworthiness reported in the literature.

The current line of inquiry could also benefit from future research in several other ways. One limitation relevant to Studies 1–3 is the use of photos of known individuals from websites. Although the majority of these photos came from official sources (e.g., company and government websites), we also relied on some media outlets for obtaining photographs. Thus, although we were careful to select photos that were created and posted prior to public report of the individuals' crimes, these faces might have been influenced by other factors. For example, the face of Ken Lay, the infamous former CEO of the Enron energy company, might have been influenced by the stress of internal turmoil within his company building up to the scandal for which he was eventually indicted, if not by the financial troubles that might have led to his criminal acts. Notably, we would have expected this to exacerbate the difference between the faces of individuals like Ken Lay and noncriminal executives, but it highlights the importance of considering the source of photographs in work on trustworthiness, as illustrated to be a critical factor in Study 1.

Fortunately, Study 4 helped to correct for this limitation in the stimuli by using photographs taken under standardized conditions in the lab. The severity of the targets' transgressions in Study 4, however, was much less than those of the targets in Studies 1–3. Although this variation across the studies is a strength of the current work, there might be other domains in which particular varieties of dishonesty may be legible from nonverbal and appearance cues, as suggested above. Specifically, might dishonest individuals be more easily detected when the stakes are lower? The rewards that might come from successfully behaving dishonestly in corporate business and in war are very high. Thus, it is possible that a great many individuals would choose to act dishonestly if placed in similar circumstances (see Zimbardo, 2007). Yet the rewards for other crimes, such as shoplifting in a supermarket, may be rather low. Thus, individuals with low thresholds for dishonest behavior may be more easily distinguished than those who would only be tempted to act dishonestly regarding matters of great importance. Although the stakes in Study 4 were arguably rather low (a $100 prize), it would be interesting to see how impressions of trustworthiness might vary for those participants who would still cheat for an even smaller $10 prize. Future work may wish to consider this type of variation as a factor.

In sum, the accuracy of judgments from appearance and nonverbal behavior may be quite domain-specific. Although individuals were able to judge targets' extraversion and intelligence from photos of their faces, their judgments of trustworthiness were not correlated with the targets' behavior. We demonstrated this in several ecologically valid domains: corporate financial scandals, war crimes, and student cheating. Data from these three areas converged to show that reliable assessments of others' trustworthiness could not be made from their faces. Moreover, the amygdala responded to perceived but not actual trustworthiness. Thus, it might be wise not to trust one's first impressions of trustworthiness.

## References

Adams, R. B., Jr., Gordon, H. L., Baird, A. A., Ambady, N., & Kleck, R. E. (2003). Effects of gaze on amygdala sensitivity to anger and fear faces. *Science, 300,* 1536. doi:10.1126/science.1082244

Adolphs, R., Baron-Cohen, S., & Tranel, D. (2002). Impaired recognition of social emotions following amygdala damage. *Journal of Cognitive Neuroscience, 14,* 1264–1274. doi:10.1162/089892902760807258

Adolphs, R., Sears, L., & Piven, J. (2001). Abnormal processing of social information from faces in autism. *Journal of Cognitive Neuroscience, 13,* 232–240. doi:10.1162/089892901564289

Adolphs, R., Tranel, D., & Damasio, A. R. (1998). The human amygdala in social judgment. *Nature, 393,* 470–474. doi:10.1038/30982

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology, 32,* 201–271. doi:10.1016/S0065-2601(00)80006-4

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7,* 268–277. doi:10.1038/nrn1884

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology, 66,* 5–20. doi:10.1037/0022-3514.66.1.5

Bishop, S., Duncan, J., Brett, M., & Lawrence, A. D. (2004). Prefrontal cortical function and anxiety: Controlling attention to threat-related stimuli. *Nature Neuroscience, 7,* 184–188. doi:10.1038/nn1173

Bond, C. F., Jr., Berry, D. S., & Omar, A. (1994). The kernel of truth in judgments of deceptiveness. *Basic and Applied Social Psychology, 15,* 523–534. doi:10.1207/s15324834basp1504_8

Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10,* 214–234. doi:10.1207/s15327957pspr1003_2

Bond, C. F., Jr., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin, 134,* 477–492. doi:10.1037/0033-2909.134.4.477

Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology, 62,* 645–657. doi:10.1037/0022-3514.62.4.645

Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *Journal of Personality and Social Psychology, 65,* 546–553. doi:10.1037/0022-3514.65.3.546

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press.

Carré, J. M., & McCormick, C. M. (2008). In your face: Facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proceedings of the Royal Society: B*. Biological Sciences, *275,* 2651–2656. doi:10.1098/rspb.2008.0873

Carré, J. M., McCormick, C. M., & Mondloch, C. J. (2009). Facial structure is a reliable cue of aggressive behavior. *Psychological Science, 20,* 1194–1198. doi:10.1111/j.1467-9280.2009.02423.x

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with Wason selection task. *Cognition, 31,* 187–276. doi:10.1016/0010-0277(89)90023-1

Cunningham, W. A., Van Bavel, J. J., & Johnsen, I. R. (2008). Affective flexibility: Evaluative processing goals shape amygdala activity. *Psychological Science, 19,* 152–160. doi:10.1111/j.1467-9280.2008.02061.x

Davis, S. F., & Ludvigson, H. W. (1995). Additional data on academic dishonesty and a proposal for remediation. *Teaching of Psychology, 22,* 119–121. doi:10.1207/s15328023top2202_6

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129,* 74–118. doi:10.1037/0033-2909.129.1.74

DePaulo, B. M., & Rosenthal, R. (1979). Telling lies. *Journal of Personality and Social Psychology, 37,* 1713–1722. doi:10.1037/0022-3514.37.10.1713

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24,* 285–290. doi:10.1037/h0033731

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience, 19,* 1508–1519. doi:10.1162/jocn.2007.19.9.1508

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102,* 652–670. doi:10.1037/0033-295X.102.4.652

Hart, A. J., Whalen, P. J., Shin, L. M., McInerney, S. C., Fischer, H., & Rauch, S. L. (2000). Differential response in the human amygdala to racial outgroup vs. ingroup face stimuli. *NeuroReport, 11,* 2351–2354. doi:10.1097/00001756-200008030-00004

Honz, K., Kiewra, K. A., & Yang, Y. (2010). Cheating perceptions and prevalence across academic settings. *Mid-Western Educational Researcher, 23,* 10–17.

Judd, C. M., Ryan, C. S., & Park, B. (1991). Accuracy in the judgment of in-group and out-group variability. *Journal of Personality and Social Psychology, 61,* 366–379. doi:10.1037/0022-3514.61.3.366

Krumhuber, E., Manstead, A. S., Cosker, D., Marshall, D., Rosin, P. L., & Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion, 7,* 730–735. doi:10.1037/1528-3542.7.4.730

Liang, X., Zebrowitz, L. A., & Zhang, Y. (2010). Neural activation in the "reward circuit" shows a nonlinear response to facial attractiveness. *Social Neuroscience, 5,* 320–334. doi:10.1080/17470911003619916

Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1995). The dissection of selection in person perception: Inhibitory processes in social stereotyping. *Journal of Personality and Social Psychology, 69,* 397–407. doi:10.1037/0022-3514.69.3.397

Malatesta, C. Z., Fiore, M. J., & Messina, J. J. (1987). Affect, personality, and facial expression characteristics of older people. *Psychology and Aging, 2,* 64–69. doi:10.1037/0882-7974.2.1.64

McArthur, L. Z., & Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological Review, 90,* 215–238. doi:10.1037/0033-295X.90.3.215

McCabe, D. L., Treveno, L. K., & Butterfield, K. D. (2001). Cheating in academic institutions: A decade of research. *Ethics & Behavior, 11,* 219–232. doi:10.1207/S15327019EB1103_2

McClelland, D. C. (1973). Testing for competence rather than for "intelligence". *American Psychologist, 28,* 1–14. doi:10.1037/h0034092

Morris, J. S., Ohman, A., & Dolan, R. J. (1998). Conscious and unconscious emotional learning in the human amygdala. *Nature, 393,* 467–470. doi:10.1038/30976

Murphy, N. A. (2007). Appearing smart: The impression management of intelligence, person perception accuracy, and behavior in social interaction. *Personality and Social Psychology Bulletin, 33,* 325–339. doi:10.1177/0146167206294871

Murphy, N. A., Hall, J. A., & Colvin, C. R. (2003). Accurate intelligence assessments in social interactions: Mediators and gender effects. *Journal of Personality, 71,* 465–493. doi:10.1111/1467-6494.7103008

Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin, 35,* 1661–1671. doi:10.1177/0146167209346309

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

O'Sullivan, M. (2008). Home runs and humbugs: Comment on Bond and DePaulo (2008). *Psychological Bulletin, 134,* 493–497. doi:10.1037/0033-2909.134.4.493

Penton-Voak, I. S., Pound, N., Little, A. C., & Perrett, D. I. (2006). Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social Cognition, 24,* 607–640. doi:10.1521/soco.2006.24.5.607

Porter, S., England, L., Juodis, M., ten Brinke, L., & Wilson, K. (2008). Is the face a window to the soul? Investigation of the accuracy of intuitive judgments of the trustworthiness of human faces. *Canadian Journal of Behavioural Science, 40,* 171–177. doi:10.1037/0008-400X.40.3.171

Rotter, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist, 35,* 1–7. doi:10.1037/0003-066X.35.1.1

Rule, N. O., & Ambady, N. (2008). The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological Science, 19,* 109–111. doi:10.1111/j.1467-9280.2008.02054.x

Rule, N. O., Ambady, N., & Adams, R. B., Jr. (2009). Personality in perspective: Judgmental consistency across orientations of the face. *Perception, 38,* 1688–1699. doi:10.1068/p6384

Rule, N. O., Ambady, N., Adams, R. B., Jr., Ozono, H., Nakashima, S., Yoshikawa, S., & Watabe, M. (2010). Polling the face: Prediction and consensus across cultures. *Journal of Personality and Social Psychology, 98,* 1–15. doi:10.1037/a0017673

Rule, N. O., Moran, J. M., Freeman, J. B., Whitfield-Gabrieli, S., Gabrieli, J. D. E., & Ambady, N. (2011). Face value: Amygdala response reflects

the validity of first impressions. *NeuroImage, 54,* 734–741. doi:10.1016/j.neuroimage.2010.07.007

Said, C. P., Baron, S., & Todorov, A. (2009). Nonlinear amygdala response to face trustworthiness: Contributions of high and low spatial frequency information. *Journal of Cognitive Neuroscience, 21,* 519–528. doi:10.1162/jocn.2009.21041

Said, C. P., Dotsch, R., & Todorov, A. (2010). The amygdala and FFA track both social and non-social face dimensions. *Neuropsychologia, 48,* 3596–3605. doi:10.1016/j.neuropsychologia.2010.08.009

Schiller, D., Freeman, J. B., Mitchell, J. P., Uleman, J. S., & Phelps, E. A. (2009). A neural mechanism of first impressions. *Nature Neuroscience, 12,* 508–514. doi:10.1038/nn.2278

Searle, W., & Ward, C. (1990). The prediction of psychological and socio-cultural adjustment during cross-cultural transitions. *International Journal of Intercultural Relations, 14,* 449–464. doi:10.1016/0147-1767(90)90030-Z

Slotnick, S. D., Moo, L. R., Segal, J. B., & Hart, J., Jr. (2003). Distinct prefrontal cortex activity associated with item memory and source memory for visual shapes. *Cognitive Brain Research, 17,* 75–82. doi:10.1016/S0926-6410(03)00082-X

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. *Psychology, Public Policy, and Law, 13,* 1–34. doi:10.1037/1076-8971.13.1.1

Stillman, T. F., Maner, J. K., & Baumeister, R. F. (2010). A thin slice of violence: Distinguishing violent from nonviolent sex offenders at a glance. *Evolution and Human Behavior, 31,* 298–303. doi:10.1016/j.evolhumbehav.2009.12.001

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

Todorov, A., Baron, S., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: A model based approach. *Social Cognitive and Affective Neuroscience, 3,* 119–127. doi:10.1093/scan/nsn009

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27,* 813–833. doi:10.1521/soco.2009.27.6.813

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

Verplaetse, J., Vanneste, S., & Braeckman, J. (2007). You can judge a book by its cover: The sequel. A kernel of truth in predictive cheating detection. *Evolution and Human Behavior, 28,* 260–271. doi:10.1016/j.evolhumbehav.2007.04.006

von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences, 34,* 1–16. doi:10.1017/S0140525X10001354

Vuilleumier, P., Armony, J. L., Clarke, K., Husain, M., Driver, J., & Dolan, R. J. (2002). Neural response to emotional faces with and without awareness: Event-related fMRI in a parietal patient with visual extinction and spatial neglect. *Neuropsychologia, 40,* 2156–2166. doi:10.1016/S0028-3932(02)00045-3

Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2003). Distinct spatial frequency sensitivities for processing faces and emotional expressions. *Nature Neuroscience, 6,* 624–631. doi:10.1038/nn1057

Weisbuch, M., Ivcevic, Z., & Ambady, N. (2009). On being liked on the web and in the "real world": Consistency in first impressions across personal webpages and spontaneous behavior. *Journal of Experimental Social Psychology, 45,* 573–576. doi:10.1016/j.jesp.2008.12.009

Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M., & Jenike, M. A. (1998). Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *The Journal of Neuroscience, 18,* 411–418.

Winston, J. S., O'Doherty, J., Kilner, J. M., Perrett, D. I., & Dolan, R. J. (2007). Brain systems for assessing facial attractiveness. *Neuropsychologia, 45,* 195–206. doi:10.1016/j.neuropsychologia.2006.05.009

Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional responses during evaluation of trustworthiness of faces. *Nature Neuroscience, 5,* 277–283. doi:10.1038/nn816

Zebrowitz, L. A. (1997). *Reading faces: Window to the soul?* Boulder, CO: Westview Press.

Zebrowitz, L. A., & Collins, M. A. (1997). Accurate social perception at zero acquaintance: The affordances of a Gibsonian approach. *Personality and Social Psychology Review, 1,* 204–223. doi:10.1207/s15327957pspr0103_2

Zebrowitz, L. A., Hall, J. A., Murphy, N. A., & Rhodes, G. (2002). Looking smart and looking good: Facial cues to intelligence and their origins. *Personality and Social Psychology Bulletin, 28,* 238–249. doi:10.1177/0146167202282009

Zebrowitz, L. A., Kikuchi, M., & Fellous, J. M. (2010). Facial resemblance to emotions: Group differences, impression effects, and race stereotypes. *Journal of Personality and Social Psychology, 98,* 175–189. doi:10.1037/a0017990

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass, 2,* 1497–1517. doi:10.1111/j.1751-9004.2008.00109.x

Zebrowitz, L. A., & Rhodes, G. (2004). Sensitivity to "bad genes" and the anomalous face overgeneralization effect: Cue validity, cue utilization, and accuracy in judging intelligence and health. *Journal of Nonverbal Behavior, 28,* 167–185. doi:10.1023/B:JONB.0000039648.30935.1b

Zebrowitz, L. A., Voinescu, L., & Collins, M. A. (1996). "Wide-eyed" and "crooked-faced": Determinants of perceived and real honesty across the life span. *Personality and Social Psychology Bulletin, 22,* 1258–1269. doi:10.1177/01461672962212006

Zimbardo, P. G. (2007). *The Lucifer effect: Understanding how good people turn evil*. New York, NY: Random House.

Received May 16, 2012
Revision received October 23, 2012
Accepted October 24, 2012 ■