Seeing is believing: observer perceptions of trait trustworthiness predict perceptions of honesty in high-stakes emotional appeals

Alysha Baker, Stephen Porter, Leanne ten Brinke & Crystal Mundy

To cite this article: Alysha Baker, Stephen Porter, Leanne ten Brinke & Crystal Mundy (2016): Seeing is believing: observer perceptions of trait trustworthiness predict perceptions of honesty in high-stakes emotional appeals, Psychology, Crime & Law, DOI: 10.1080/1068316X.2016.1190844

To link to this article: http://dx.doi.org/10.1080/1068316X.2016.1190844
Seeing is believing: observer perceptions of trait trustworthiness predict perceptions of honesty in high-stakes emotional appeals

Alysha Baker\textsuperscript{a}, Stephen Porter\textsuperscript{a}, Leanne ten Brinke\textsuperscript{b} and Crystal Mundy\textsuperscript{a}

\textsuperscript{a}Centre for the Advancement of Psychological Science and Law (CAPSL), University of British Columbia, Kelowna, Canada; \textsuperscript{b}Haas School of Business, University of California Berkeley, Berkeley, CA, USA

\textbf{ABSTRACT}
Instantaneous first impressions of facial trustworthiness influence the manner in which observers evaluate ensuing information about stranger targets [e.g. Porter, S., & ten Brinke, L. (2009). Dangerous decisions: A theoretical framework for understanding how judges assess credibility in the courtroom. \textit{Legal and Criminological Psychology}, 14, 119–134. doi:10.1348/135532508X281520]. In two studies, we examined the association between perceptions of general trustworthiness and honesty assessments in an extremely high-stakes sample – individuals publicly pleading for the return of a missing relative, half of whom had killed the missing individual. In Study 1, observers ($N = 131$) provided trustworthiness ratings – either before or after viewing and evaluating the honesty of videotaped or audio-only pleas – for a still image that depicted a neutral expression on the face of each pleader. In Study 2, observers ($N = 220$) evaluated the sincerity of audio pleas paired either with an untrustworthy-looking target, a trustworthy-looking target, or no target face. Collectively, our findings indicated that first impressions of trait trustworthiness form the basis of state judgments of honesty, potentially contributing to misguided credibility assessments and miscarriages of justice in the legal system.

Upon encountering a stranger, observers rapidly form inferences about that person’s state and trait characteristics (e.g. Martelli, Majaj, & Pelli, 2005; Olivola & Todorov, 2010; Rule & Ambady, 2010). In the absence of other information, observers rely on ‘intuitive’ judgments and heuristics to inform the course of ensuing actions. Evaluations of trait trustworthiness, that is, the basic discrimination of friend and foe, likely was one of the earliest interpersonal judgments to evolve (e.g. Williams & Mattingley, 2006) and is one of the most expedient and enduring interpersonal assessments; inferences of trustworthiness, based on a stranger's face, occur within 38 ms, and while confidence in judgment accuracy increases with longer exposure to the face, initial assessments remain virtually unchanged (Bar, Neta, & Linz, 2006; Willis & Todorov, 2006).
First impressions of trait trustworthiness

Both dynamic emotional expression and relatively unchanging aspects of facial structure predict who is deemed (un)trustworthy. For example, a stranger approaching with flared nostrils, gritted teeth, and furrowed brows exhibits anger or rage, reliably suggests a potential threat, and may be evaluated as untrustworthy and scary (e.g. Marsh, Ambady, & Kleck, 2005). The evaluation of emotional expressions as a means of identifying trustworthiness, however, appears to have overgeneralized. That is, facial structures that resemble particular emotions lead to similar conclusions; for example, an individual whose facial structure includes downward angled eyebrows may be misperceived as angry or untrustworthy (Oosterhof & Todorov, 2009). Similarly, structural characteristics such as higher eyebrows, more pronounced cheekbones, and wider chins signal trustworthiness, as they more closely resemble displays of happiness and surprise (Bar et al., 2006; Todorov, 2008; Todorov, Baron, & Oosterhof, 2008; Vartanian et al., 2012; Willis & Todorov, 2006). Attending to structural features indicative of dominance may further contribute to an evaluation of potential threat. Specifically, a more dominant-looking (testosterone-laden) individual may be evaluated as untrustworthy because of this perceived strength and the consequences of his potential aggression. Indeed, Oosterhof and Todorov (2008) found that valence and dominance are important contributors to perceptions of trustworthiness, and that as faces are structurally manipulated to appear less trustworthy, they are perceived as expressing greater hostility. In short, a target’s facial appearance leads to the formation of observer impressions of trait trustworthiness.

This first impression formation occurs so rapidly that it is unlikely to be a conscious process (e.g. ‘He has downturned eyebrows, therefore I do not trust him’) but rather may be experienced as ‘intuition’ about the target (e.g. ten Brinke & Porter, 2011). This type of intuition is encouraged throughout society. For example, a mother may tell her daughter to ‘trust your instinct’ around men. In the legal system, judges encourage jurors to use their intuition in evaluating the credibility of witnesses (e.g. R. v. Lifchus, 1997). However, intuitive judgments of trustworthiness appear to be unreliable; for example, Rule, Krendl, Ivcevic, and Ambady (2013) found that inferences of trustworthiness attributed to the faces of corporate criminals were equivalent to those of noncriminal executives and, similarly, that the perceived trustworthiness of military criminals and military heroes did not differ. Other research has found that observers can only differentiate between Nobel Peace Prize recipients/humanitarians and America’s Most Wanted criminals at slightly above chance (Porter, England, Juodis, ten Brinke, & Wilson, 2008). Thus, in the modern day, such evaluations may play a small or no facilitative role in forming accurate impressions of trustworthiness.

Dangerous decisions: how trait trustworthiness affects legal decision-making

Despite the unreliability of trustworthiness judgments, they appear to influence judgments about others’ intentions, abilities, and anticipated behavior in various contexts. For example, research has highlighted that mate selection (e.g. Olivola et al., 2009), election results (e.g. Olivola & Todorov, 2010; Todorov, Mandisodza, Goren, & Hall, 2005), and hiring decisions may be strongly influenced by first impressions of trustworthiness (e.g.
Ambady & Rosenthal, 1992). The Dangerous Decisions Theory (DDT; Porter & ten Brinke, 2009) predicts that trait trustworthiness assessments have an enduring influence on the manner in which new information concerning the target is interpreted by observers. Using a courtroom example, the initial intuitive evaluation of a defendant’s trait trustworthiness by a judge or juror will influence subsequent inferences concerning the defendant’s credibility while on the stand, generating a non-critical, ‘tunnel vision’ assimilation of potentially ambiguous or contradictory evidence concerning the defendant. Accordingly, defendants perceived to be untrustworthy-looking are more likely to receive guilty verdicts from mock jury participants for serious crimes based on less evidence than their trustworthy-looking counterparts (e.g. Porter, ten Brinke, & Gustaw, 2010). Recently, Wilson and Rule (2015) found that untrustworthy-looking men convicted of murder in the USA were more likely than trustworthy-looking men convicted of similar crimes to be handed the death penalty instead of life in prison.

Despite support for the impact of first impressions of trustworthiness on legal decision-making, it has yet to be directly established that perceived trustworthiness biases perceptions of honesty, specifically. We propose that impressions of trait trustworthiness may contribute to the low accuracy rates typically associated with detecting deception (e.g. Bond & DePaulo, 2006; Vrij, Granhag, & Porter, 2010). In other words, inaccurate impressions of trustworthiness — guided by intuitive assessments of facial features — may bias assessments of state honesty such that trustworthy faces are more often thought to be telling the truth, and untrustworthy faces to be lying, regardless of their actual veracity.

**High-stakes emotional deception**

Although Hartwig and Bond (2011) suggest that poor deception detection accuracy is the result of weakness of cues to deception (e.g. Vrij et al., 2010), such behavioral indicators are theorized to be more salient and reliable in high-stakes situations (i.e. motivational impairment effect; DePaulo & Kirkendol, 1989; but see Hartwig & Bond, 2014). Research suggests that emotional facial expressions and verbal cues can be powerful indicators of high-stakes deception (Archer & Lansley, 2015; McQuaid, Woodworth, Hutton, Porter, & ten Brinke, 2015; ten Brinke & Porter, 2012; Wright Whelan, Wagstaff, & Wheatcroft, 2014, 2015). While some research suggests that these high-stakes lies in particular are detected above chance (e.g. Wright Whelan et al., 2015), other research has not found support for this conclusion (e.g. Baker, ten Brinke, & Porter, 2013; Evanoff, Porter, & Black, 2014; Vrij & Mann, 2001). In one investigation, even police officers were unable to discriminate genuine versus deceptive pleaders above the level of chance (Vrij & Mann, 2001). Regardless of the level of accuracy that can be achieved by attending to valid behavioral cues in these high-stakes emotional pleas, the DDT model would suggest that high-stakes environments actually may amplify reliance on erroneous heuristics and tunnel visions, including biased conclusions based on the face (Porter & ten Brinke, 2009). Thus, while reliable cues may be available, the influence of first impressions may prevail.

**The logical leap from trait trustworthiness to state honesty**

To our knowledge, only one previous study has directly examined the relation between evaluations of trait trustworthiness and state honesty. O’Sullivan (2003) found that
judgments of target trustworthiness were correlated with state judgments of truthfulness in the predicted manner (e.g. a target who is judged trustworthy-looking is more likely to be classified as telling the truth). In the study, observers evaluated the sincerity of two types of stories – strong opinions (e.g. about capital punishment) and mock thefts. Observers of the first sample of 10 targets initially evaluated whether each target had been lying and then rated his/her trustworthiness; however, an important limitation of this approach is that the explicit honesty evaluations could have biased the subsequent trustworthiness judgments. Observers of the second sample of 10 targets evaluated trustworthiness from a still-video frame and then evaluated the honesty of the theft denials. While this approach partially addressed the limitation of part one of the study, the still frames were not controlled for facial expression (so, for example, one frame could depict an anger expression while another could depict a smiling target). Further, another critical limitation is that these studies did not examine the alternative possibility that untrustworthy-looking individuals actually provide less credible stories. This alternative explanation could have been addressed by including audio-only (control) conditions (i.e. making trustworthiness evaluations of a group of still frames and then evaluating the honesty of the audio-only stories and, alternatively, evaluating the honesty of the audio-only stories and then evaluating the trustworthiness of the group of still frames) to ensure that untrustworthy-looking individuals do not simply provide less credible verbal accounts.

The present study

In two studies, we sought to address the limitations of past research and to determine whether there was evidence for first impressions of trait trustworthiness guiding judgments of state honesty using a high-stakes forensic sample, which, previous research suggests, would be most likely to elicit the hypothesized effect (Porter et al., 2010; Wilson & Rule, 2015). In Study 1, still images of individuals giving a public televised plea for the safe return of a missing family member were presented (along with prompts for trustworthiness ratings) either prior to or after making an honesty evaluation based on the videotaped plea. Specifically, it was expected that observers’ perceptions of pleader trustworthiness would be related to increased evaluations of honesty. However, it was predicted that there would be no relation between perceived trustworthiness and actual veracity because recent research highlights the lack of relation between perceived and actual (behavioral) trustworthiness (e.g. Rule et al., 2013) and, secondly, observers typically perform around the level of (or slightly above) chance in detecting lies (e.g. Bond & DePaulo, 2006). In contrast to the expected relation between perceived trustworthiness and honesty assessments in the video conditions, we did not expect this effect in audio conditions in which facial appearance was unavailable to bias honesty assessments. Such a result would also suggest that O’Sullivan’s (2003) findings cannot be explained by trustworthy-looking liars providing more compelling verbal statements.

Study 2 utilized an experimental design to further address the current research question. Importantly, observers provided honesty assessments on faces pre-rated on trustworthiness by a pilot sample; as such, there would be no opportunity for the influence of consistency responding on observers’ decisions (i.e. rating a target as honest because he/she had previously rated him/her as trustworthy, or vice versa). Here, audio-only pleas were paired with a trustworthy-looking face, an untrustworthy-looking face, or no
face to examine veracity judgments. It was hypothesized that pleas paired with trustworthy faces would be more likely to be deemed truthful, whereas pleas paired with untrustworthy faces would be deemed deceptive. These hypothesized results would provide additional support for the significant weight given to rapid inferences of trustworthiness from a glimpse of a target’s face and the manner in which these evaluations influence later decision-making with a unique real-world target sample of genuine intensely distressed family pleaders versus deceptive killers engaging in highly motivated acting.

Study 1

Methods

Participants
Undergraduate participants (N = 131) attending a Canadian university were recruited through an online research participant pool and received course credit for completion of the study. The sample consisted of 90 women and 41 men with a mean age of 20.58 years (SD = 2.98).

Materials
Pleas. Stratified random sampling (veracity, gender) was used to select 20 videos of individuals pleading for the safe return of a missing family member from a larger sample (see ten Brinke & Porter, 2012). Videos were collected from news agencies in Australia, Canada, the United Kingdom, and the United States. Each videotaped individual made a direct appeal to the (supposed) perpetrator to release the missing person, the missing person to make contact, or to the public for information/search party assistance. Of the 20 videos, 10 individuals (5 deceptive males and 5 deceptive females) eventually had been convicted of murdering the missing person based on overwhelming evidence (e.g. DNA). In the case of genuine pleaders, 10 individuals (5 males and 5 females) were innocent of any involvement in the missing person case – either another person was convicted of murdering the missing person based on similarly overwhelming evidence, or the missing individual was later located in the absence of foul play. For more information on the nature of the videos and the determination of ground truth in these cases, see ten Brinke and Porter (2012). The videos were created from their original long version to include only the direct plea in which the pleader spoke directly to (a) the missing person, asking him/her to come home, (b) the perpetrator, asking him/her to let the missing person go, and/or (c) the public, asking them to join in the search for the missing individual. Videos ranged from 8 to 27 s, with an average length of 17 s.

Still images. Still images were selected to assess first impressions based on the facial appearance. A still image was selected from a neutral facial expression present at some point during each plea, according to the coding scheme developed by Porter and ten Brinke (2008; also see ten Brinke, Porter, & Baker, 2012). Two trained coders (who previously had established high reliability in coding emotional expressions according to this method) agreed that selected faces expressed no emotion in 100% of cases (α = 1.00).
Procedure

Each participant was provided a website address that he/she could follow to complete the study online. Participants were randomly assigned to one of four conditions that varied in task order and stimulus presentation. In the still-video condition (n = 35), participants first viewed a group of still images (one for each pleader) in randomized order, for as much time as was needed, and provided trustworthiness ratings, ranging from 1 (not at all) to 7 (extremely). Participants pressed ‘next’ to proceed to the next face until all 20 were rated. After a delay task (during which individual difference measures were completed), participants viewed the videotaped (with sound) pleas in randomized order and provided dichotomous honesty judgments (i.e. genuine or deceptive). In other words, the dichotomous honesty judgment occurred directly after the corresponding video was viewed. Presentation of still images and videotaped pleas was counterbalanced to examine order effects and the delay task was introduced to reduce the chance of conscious attempts to respond consistently across tasks. In the video-still condition (n = 24), participants began the study by viewing and evaluating the pleas in video (with sound) format in randomized order and subsequently (following completion of a delay task during which individual difference measures were completed) viewed (for as much time as was needed) and evaluated the trustworthiness of the group of still images (of the same targets) in randomized order. These conditions were similar to those used in the O’Sullivan (2003) study and would serve to examine the magnitude of effects for the explicit trustworthiness evaluations versus honesty assessments, on the other (subsequently completed) task.

In the still-audio condition (n = 38), participants completed the same tasks as in the still-video condition (described above) but instead of viewing an audio-visual version of the plea, they listened to an audio-only version. That is, they rated the trustworthiness of all 20 still images in random order, completed the delay task, and then listened to and evaluated the honesty of their pleas in random order. As such, participants in these audio-only conditions would be unable to associate particular target faces with a particular target story/plea. Finally, the audio-still condition (n = 34) was identical to the video-still condition but participants instead began the study by listening to the pleas (and evaluating honesty) in audio-only format in random order, completed a delay task, and then viewed and evaluated the trustworthiness of a group of still frame images. The two audio conditions served as control conditions to rule out the alternative explanation that untrustworthy-looking individuals simply provide less convincing pleas.

Results

Relation between trustworthiness and honesty assessments. The data first were transposed so that the relationship between trustworthiness ratings and honesty judgments could be examined with the pleader as the unit of analysis, rather than the individual participant. As such, mean trustworthiness ratings for the still image and percentage of honesty ratings were calculated for each of the 20 faces. See Table 1 for basic descriptive statistics associated with each.

In the still-video condition, in which participants provided honesty assessments following their trustworthiness ratings, trustworthiness ratings were significantly positively related to the percentage of participants who judged the pleader to be honest (r(18)}
Similarly, in the video-still (i.e. reverse order) condition, in which participants provided trustworthiness ratings following their honesty assessment, trustworthiness ratings were strongly positively correlated to the percentage of honesty judgments \( r(18) = .61, p = .004 \). However, as expected, there were no significant relationships identified in either the still-audio \( (r(18) = .22, p = .35) \) or audio-still \( (r(18) = -.12, p = .65) \) condition, suggesting that untrustworthy-looking pleaders did not simply provide less convincing pleas. Further, trustworthiness ratings associated with the still-video \( (r(18) = -.07, p = .77) \), video-still \( (r(18) = .06, p = .80) \), still-audio \( (r(18) = -.02, p = .95) \), or audio-still \( (r(18) = .12, p = .65) \) were not significantly related to the number of individuals who accurately classified each pleader as honest or deceptive.

The overall accuracy rate \( (M = 50.38\%, SD = 17.73) \) at detecting veracity was not significantly different from chance, \( t(79) = 0.19, p = .85, d = 0.04 \). Participants in the video versus audio conditions did not perform differently from each other, \( t(78) = 0.16, p = .87, d = 0.04 \), and neither group exceeded chance. Video percentage accuracy and audio percentage accuracy were not significantly different from chance either, \( t(39) = 0.24, p = .81, d = 0.08 \), and \( t(39) = 0.02, p = .98, d = 0.006 \), respectively. Further, no gender differences in accuracy were evidenced overall, or within audio and video conditions, \( ps > .05 \).

**Comparing honesty decisions across conditions.** In order to compare the four conditions and gain a better understanding of the magnitude of the previously reported relationships, a correlation between the 20 trustworthiness ratings and corresponding honesty assessments was created for each participant in the study. In other words, for a given participant, a correlation was calculated between his/her trustworthiness ratings and honesty assessments. This procedure was repeated for each participant generating a variable indicating the extent to which trustworthiness ratings and honesty assessments were related, at the participant level. The magnitude of these relationships was compared across condition (still-video, video-still, still-audio, audio-still) using a one-way ANOVA. A significant main effect was evidenced, \( F(3, 121) = 16.48, p < .001, \eta^2 = .30 \) (see Table 2 for descriptive statistics). A series of independent-sample \( t \)-tests was conducted comparing the correlations from video conditions to audio conditions and the two video conditions to each

| Table 1. Descriptive statistics for trustworthiness ratings, percentage honesty assessments, and accuracy scores (study 1). |
|-----------------|-----------------|-----------------|
| Trustworthiness ratings (M, SD) | Percentage honesty assessments (M, SD) | Accuracy (%) |
| Still-video condition | 3.36 (0.81) | 63.00 (13.26) | 50.99 (19.36) |
| Video-still condition | 3.54 (0.73) | 60.41 (14.27) | 50.41 (17.82) |
| Still-audio condition | 3.37 (0.73) | 60.34 (10.90) | 50.28 (15.21) |
| Audio-still condition | 3.51 (0.74) | 63.67 (13.65) | 49.86 (19.56) |

The procedure was repeated for each participant generating a variable indicating the extent to which trustworthiness ratings and honesty assessments were related, at the participant level. The magnitude of these relationships was compared across condition (still-video, video-still, still-audio, audio-still) using a one-way ANOVA. A significant main effect was evidenced, \( F(3, 121) = 16.48, p < .001, \eta^2 = .30 \) (see Table 2 for descriptive statistics). A series of independent-sample \( t \)-tests was conducted comparing the correlations from video conditions to audio conditions and the two video conditions to each

| Table 2. Descriptive statistics for participant-level correlations for each presentation condition (study 1). |
|-----------------|-----------------|-----------------|
|                  | M               | SD              |
| Still-video condition | 0.14            | 0.24            |
| Video-still condition | 0.41            | 0.24            |
| Still-audio condition | 0.06            | 0.21            |
| Audio-still condition | -0.01           | 0.24            |

---

*Downloaded by [128.189.114.181] at 07:13 20 July 2016*
other. Mean correlations associated with audio conditions ($M = 0.03, SD = 0.23$) and video conditions ($M = 0.25, SD = 0.27$) were significantly different, $t(120) = 5.05, p < .001, d = 0.92$. The magnitude of the relationship was significantly lower in the still-video ($M = 0.14, SD = 0.24$) relative to the video-still ($M = 0.41, SD = 0.24$) condition, $t(55) = 4.22, p < .001, d = 1.13$. Further, the magnitude of the relationship was not significantly different between the still-audio ($M = 0.06, SD = 0.21$) and audio-still condition ($M = -0.01, SD = 0.24$), $t(63) = -1.25, p = .22, d = 0.31$. Additional analyses revealed that the video-still and audio-still conditions were significantly different from each other, $t(53) = 6.52, p < .001, d = 1.75$. However, while we would expect the magnitude of the relationships between facial trustworthiness ratings and honesty assessments to be greater in the still-video versus still-audio condition, the two conditions did not reach statistical significance, $t(65) = 1.52, p = .13, d = 0.36$.

A series of follow-up one-sample $t$-tests was conducted to examine whether the mean correlations between honesty judgments and trustworthiness ratings for each condition differed from a value of 0 (indicating no relationship between the two judgments). Analyses revealed that mean correlations between judgments in the still-video condition were significantly different from 0, $t(33) = 3.56, p = .001, d = 1.24$. Furthermore, the mean correlations between judgments in the video-still condition also were significantly different from 0, $t(23) = 8.40, p < .001, d = 3.50$. The still-audio and audio-still conditions, however, were not significantly different from 0, $t(32) = 1.60, p = .12, d = 0.57$ and $t(32) = -0.24, p = .81, d = 0.09$, respectively. Although the earlier analyses indicated that still-video and still-audio relationships did not differ significantly, these analyses reveal that the still-video, and not the still-audio condition, had a relationship magnitude greater than 0. Collectively, these findings lend further support to the conclusion of a relationship between trustworthiness and honesty evaluations in video conditions but no relationship in audio-only conditions.

**Study 2**

A second, experimental study was conducted to overcome a limitation associated with Study 1. Observers may have been making an effort to remain consistent in their honesty judgments of the pleas and trustworthiness ratings of the still images; for example, in Study 1 those deemed deceptive murderers were more likely to be classified as untrustworthy. By using an experimental design that utilizes faces rated on trustworthiness by a separate sample, we are able to overcome this concern. Further, an experimental study allows us to rule out a second alternative explanation for findings in Study 1; that is, untrustworthy-looking pleaders may be behaving more suspiciously than trustworthy-looking pleaders. We are better able to isolate the effect of facial trustworthiness by employing an experimental manipulation.

**Methods**

**Participants**

Undergraduate participants ($N = 220$) attending a Canadian university were recruited through an online research participant pool and received course credit for completion of the study. The sample consisted of 162 women and 51 men (7 choosing not to
report) with a mean age of 20.39 years (SD = 3.06). Data from participants reporting issues with the media were removed from analyses. Further, those who did not correctly answer at least three out of four attention check questions were removed from analyses, leaving a total sample of $N = 172$.

**Materials**

**Pleas.** The audio was extracted from the pleader videos used in Study 1 and these audio-only pleas were used in Study 2 by pairing them with target face conditions (i.e. the audio clips always accompanied a target face unless assigned to the no face condition; see subsequent sections).

**Target faces.** FaceGen Modeller (Version 3.2) software was used to randomly generate a face for each plea with matched gender and age to the original pleader. Next, a trustworthy-looking and untrustworthy-looking version of each randomly generated target face was created by enhancing facial features associated with trustworthiness (e.g. increase height of eyebrows and chin width) and by modifying facial features to decrease perceived trustworthiness, respectively. A separate sample of undergraduate participants ($N = 92$) were recruited to rate each of these faces, to serve as a manipulation check, and provide evidence that the computerized faces did differ on the expected dimension. As anticipated, the trustworthy- and untrustworthy-looking version of each of the 20 pleaders was rated as significantly different from each other, $p < .01$.

**Procedure**

Each participant was provided a website address that he/she could follow to complete the study online. Each participant listened to all 20 pleas and each plea was presented with a trustworthy or untrustworthy version of the target (manipulated using FaceGen Software), or no face. The face condition was randomly assigned for each plea. In other words, participants listened to all pleas, and were exposed to one presentation condition per plea (but the design was not fully within-subjects). On average, participants saw 6.76 (SD = 2.15) pleas with a trustworthy face, 6.57 (SD = 2.13) pleas with an untrustworthy face, and 6.68 (SD = 2.15) pleas with no face. Upon viewing each face and paired audio clip, participants provided a dichotomous honesty judgment (i.e. genuine or deceptive) and a continuous rating, ranging from 1 (definitely deceptive) to 7 (definitely honest). Further, participants rated their confidence in the decision, ranging from 1 (not at all) to 7 (extremely). Participants pressed ‘next’ to proceed to the next plea until all 20 were rated (in randomized order), at which point individual difference measures were completed prior to being debriefed.

**Results**

**Descriptive statistics.** A total of 3160 observations were provided by the 172 participants included in the analyses below. Of the 3160 total observations, 1062 were in the low trust condition, 1072 were in the high trust condition, and 1026 were in the no face, control condition. The overall mean veracity rating was 3.49 out of 7 (SD = 1.32) and participants labeled pleaders as honest, in their dichotomous veracity judgments, 61.65% of the time. Finally, the overall mean confidence rating for these judgments was 4.64 out of 7 (SD = 1.28).
**Examining the influence of face versions on veracity ratings.** Fixed effects regression models were used to examine the potential influence of face versions on the dependent variables of interest. In each of the models conducted here (unless specified otherwise), we used dummy variables identifying trustworthy-looking faces and untrustworthy-looking faces, therefore comparing each of these groups to the control, no face condition. To account for the potential effects of pleader gender and actual veracity (i.e. whether the pleader was deceptive or genuine during his/her plea based on actual case evidence), these variables were entered as covariates. Further, variables identifying the pleader and participant were included as fixed effects to account for face-specific and participant-specific effects.

The first fixed effects regression model was conducted with continuous honesty ratings as the dependent variable. The overall model was significant, $F(192, 2967) = 7.19$, $p < .0001$, $R^2 = .32$, $R^2_{adj} = .27$, and there was a significant effect of face condition, as evidenced by a significant Wald’s test, $F(2, 2967) = 5.34$, $p = .005$. Using the no face condition as the comparison group, results revealed that pleas accompanied by untrustworthy-looking targets were rated as more deceptive than pleas accompanied by no face, $\beta = -.17$, SE = .05, $p = .001$. However, there was no difference in veracity ratings for pleas with trustworthy-looking targets versus face-absent pleas, $p = .09$. Despite the prediction that untrustworthy-looking targets would be perceived as less honest than trustworthy-looking targets, this effect was not statistically significant, $\beta = -.08$, SE = .05, $p = .116$. Actual veracity, however, did have a significant effect in the model, $\beta = -.26$, SE = .13, $p = .04$, such that deceptive murderers were inaccurately perceived as being more honest than genuinely distressed individuals. A second fixed effects regression model was conducted with the dichotomous honest judgments and confidence in those judgments, which revealed similar findings. Coefficients for each fixed effects model and corresponding variables can be found in Table 3.

Because the relationship between trustworthy and untrustworthy face conditions did not yield a significant difference as predicted, follow-up analyses were conducted to determine whether the faces deemed untrustworthy and trustworthy for the main study were significantly different from the center point of the trustworthiness scale (i.e. ‘somewhat trustworthy’ = 4). Several one-sample t-tests comparing the selected targets to a value of 4 indicated that although the trustworthiness ratings for each of our untrustworthy faces were significantly less than 4 ($ps < .01$), the ratings given to the trustworthy faces were not always significantly greater than 4. To us, this suggested that the evidenced non-significant difference between the untrustworthy and trustworthy face conditions may be attributable to the computer-generated nature of the target faces, such that it was difficult to create highly trustworthy-looking faces using the computer program. To examine this explanation statistically, a sub-sample of the target faces were selected for comparison. Pairs of faces were retained if the corresponding trustworthy face received a mean trustworthiness rating greater than 4. Using this criterion, eight pairs of faces were removed from analysis, and our sample dropped to 1952 observations. Following this deletion, the fixed effects regression model was conducted again on the veracity judgments. Once these pairs of targets were removed, the overall model was still significant, $F(184, 1767) = 4.27$, $p < .001$, $R^2 = .31$, $R^2_{adj} = .24$, and there was a significant effect of face condition, as evidenced by a significant Wald’s test, $F(2, 1767) = 4.44$, $p = .01$. More
importantly, when using the trustworthy face condition as the comparison group, results revealed that the untrustworthy face condition was rated as significantly less honest (and more deceptive), $\beta = -0.16$, SE = .07, $p = .017$ (see Table 3 for coefficients for this restricted sample fixed effects model).

**Discussion**

The current pair of studies examined the relation between perceived trait trustworthiness based on facial appearance and state honesty assessments using a forensic sample of targets relating extremely high-stakes pleas to the public. The results indicate that initial impressions based only on facial appearance (i.e. not verbal or behavioral cues) were strongly associated with subsequent ratings of honesty. In other words, targets perceived as ‘trustworthy’ were more likely to be believed (and vice versa) during their plea for the safe return of a missing family member. Further, explicit honesty evaluations of the targets were even more strongly ($r = .69$) related to subsequent assessments of how trustworthy the targets appeared in a still frame. As such, the effect is evidenced in both directions with trait impressions influencing state judgments and vice versa. Although this effect may be interpreted as observers’ desire to provide consistent responses to related questions, a subsequent experiment found that the biasing effect persists when observers only provided honesty assessments of pre-rated untrustworthy and trustworthy faces produced by computer manipulation.

While these results corroborated those of O’Sullivan’s (2003), the effects were stronger here suggesting that the real-life, emotional, high-stakes nature of the stories may have increased motivation and involvement among our observers, exacerbating their decision-making biases. Initially, we did not anticipate a stronger effect of honesty evaluations influencing trustworthiness evaluations than vice versa (i.e. the video-still condition being greater than the still-video condition). The increased magnitude of the effect in this direction may have been evidenced because of the powerful emotional content of the pleas themselves. For example, having decided based on reviewing the emotional video...
that someone had lied and had killed their missing child may lead an observer to a powerful inference of trait untrustworthiness. Further, it may have been evidenced due to a circular and reinforcing ‘dangerous decisions’ process, such that observers in this condition (i.e. those that conducted honesty evaluation first) automatically (but not explicitly) formed a first impression of trustworthiness based on the first clip of the video which subsequently influenced his/her honesty judgment and then further biased his/her trustworthiness evaluation; however, Study 1 alone is unable to confirm this. An alternative explanation is that untrustworthy-looking pleaders behaved more suspiciously than trustworthy-looking pleaders. Although this alternative explanation cannot be ruled out by Study 1, this was another motivating factor for examining the relationship in a more controlled manner in Study 2 which allowed for an investigation of the isolated effect of facial trustworthiness. The results of Study 2 further supported the conclusion that a biasing effect is present and is in line with DDT, which suggests that first impressions are held onto more strongly over time and may be reinforced as more and more judgments are formulated (Porter & ten Brinke, 2009).

Overcoming limitations of previous research, audio-only pleas were used to clarify the specific role of the face in the relation between trustworthiness evaluations and honesty assessments. In contrast to the expected relation between trustworthiness and honesty in the video conditions, we did not anticipate this effect in audio conditions in which facial appearance was unavailable to bias credibility assessments. Although the relationship between trustworthiness and honesty assessments in the still-video and still-audio condition did not differ, as predicted, relationships between trustworthiness and honesty assessments were stronger in the video than audio conditions, and significantly greater than zero in the video conditions only. That relationships in the audio conditions did not exceed zero suggests that untrustworthy-looking pleaders did not provide less credible pleas than relatively trustworthy-looking pleaders.

Another novel aspect of the present research was that target faces were manipulated to examine the direct role of facial trustworthiness on honesty judgments (Study 2). Interestingly, only untrustworthy-looking targets were rated as deceptive more often compared to when no face was presented. That is, trustworthy-looking targets received the same judgments as when no face accompanied the plea. Although it was originally anticipated that both trustworthy- and untrustworthy-looking targets would influence honesty judgments (in opposite directions), pleas that were accompanied with no face may simply be judged in line with our standard truth bias (DePaulo, Charlton, Cooper, Lindsay, & Muhlenbruck, 1997; see also Vrij, 2008).

Further, trustworthy- and untrustworthy-looking faces did not receive significantly different honesty judgments in our initial analysis. Follow-up analyses supported the explanation that this may have resulted from relatively low trustworthiness ratings given to the targets in our trustworthy face condition, which could be attributed to the nature of our computer-generated faces. Indeed, after pairs of targets with trustworthy faces that had particularly low ratings were removed (i.e. those that fell below the center point of the trustworthiness rating scale), similar results from the main analysis were maintained and – importantly – the two target conditions varied in line with our predictions; that is, the untrustworthy face condition was rated as more deceptive than the trustworthy condition.
Collectively, the pattern from both studies further supports the DDT and builds on earlier work showing that facial trustworthiness has a strong influence on varied interpersonal decisions, including dramatic effects upon criminal sentencing (Wilson & Rule, 2015). Importantly, these findings suggest that this effect persists even in the presence of reliable behavioral deception cues that observers could have utilized to inform their honesty assessments (Hartwig & Bond, 2011; ten Brinke & Porter, 2012). The present findings suggest that first impressions could have a strong influence on subsequent decision-making in the courtroom, possibly resulting in outcomes such as wrongful convictions or acquittals. Further, the findings from Study 1 showed support for a circular and reinforcing model, such that information encountered post first impression formation influences observers’ perceptions of a target’s trustworthiness and subsequent decision-making as well as vice versa, similar to a recently found pattern in relation to memory (Baker et al., 2012).

Despite these advances, a number of limitations of the current study should be noted as avenues for future research. Undergraduate students were used as participants in this study and future research might examine the effect among legal decision-makers. Future research should also consider interventions that may be effective in reducing this bias. For example, no research has yet explored whether simply warning observers about the potential biasing effect of facial trustworthiness might reduce the magnitude of the bias. Future research should also investigate these findings with non-computer-generated targets (to overcome the issue of low trustworthiness ratings for faces manipulated to be trustworthy-looking), and should attempt to utilize a greater number of target faces to gain additional confidence in our reported effects (see Westfall, Kenny, & Judd, 2014).

In conclusion, this work provides evidence that when it comes to assessing the veracity of high-stakes targets seeing really is believing. Initial impressions of trait trustworthiness influence whether the target is believed when relating their distress in a video. While our primary interest is in how these perceptual biases may play out in a police investigation or courtroom, we suspect that they may generalize to non-forensic, everyday life (social relationships, business, and politics) contexts in which observers interact with strangers and make important inferences of trustworthiness and honesty.

Notes

1. While randomization to condition was successful, a disproportionate number of participants did not complete the video-still condition. These individuals were dropped from analyses, leaving this condition with a somewhat smaller \( n \) than the others.
2. Female observers (\( M = 3.56, SD = 0.60 \)) gave significantly higher trustworthiness ratings than male observers overall (\( M = 3.19, S = 0.92 \)), \( t(129) = -2.33, p = .02, d = 0.48 \).
3. The positive relationship between trustworthiness ratings and impressions of honesty was even evidenced in a between-subjects analysis in which observer ratings were uninfluenced by the other task. That is, ratings of facial trustworthiness from the still-audio condition were positively associated with honesty ratings in the video-still condition, \( r (18) = .40, p_{\text{one-tailed}} = .04 \).

Disclosure statement

No potential conflict of interest was reported by the authors.
Funding information

This work was supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) through a grant to the second author [grant number 410-2010-0034].

References


