AN ABSTRACT OF THE DISSERTATION OF

Zoe M. Alley for the degree of Doctor of Philosophy in Psychology presented on November 18, 2020.

Title: Implications of Person Perception Across Development: The Reciprocal Influences of Problem Behavior and Facial Trustworthiness.

Abstract approved:

_____

David C. R. Kerr

People of all ages and ethnicities implicitly use others' facial characteristics to evaluate their personalities. The field of person perception has identified several mechanisms through which one's facial appearance may be associated with one's behavior. For example, a person with an untrustworthy-looking face may elicit negative reactions from social partners, which may then cause the person to engage in more delinquency than they otherwise would have (expectancy effects), with negative outcomes for both the actor and those affected by their delinquent behavior. Alternatively, engagement in delinquency may cause a person to develop an untrustworthy appearance (a Dorian Gray effect). Such degradations in facial trustworthiness may in combination with expectancy effects interfere with desistence of delinquency during early adulthood or disadvantage persons who have desisted antisocial behavior. Thus, it is paramount to understand and interrupt both processes across development in order to reduce incidence of delinquency and encourage

desistence. Yet the investigation of facial trustworthiness has rarely been generalized to a developmental context. The present project examined both the target of interpersonal perceptions (Study 1) and the processes that lead perceivers to behave differently toward those targets (Study 2). Study 1 leveraged methods from developmental psychology to follow a sample of 206 at risk boys from ages 13 to 38. This was the first study to chart the development of facial trustworthiness across adolescence and into adulthood. Initial levels of facial trustworthiness at age 13 predicted slower escalation in delinquency during adolescence and faster declines in delinquency during adulthood (expectancy effects), and initial levels of delinquency at age 13 predicted more rapid degradations in facial trustworthiness across adolescence (Dorian Gray effects). Study 2 utilized methods from experimental psychology to investigate the extent to which ambiguous behavioral information may intensify the effect of facial trustworthiness on perceivers' evaluations, a process that may contribute to expectancy effects. However, Study 2 failed to replicate an effect of facial trustworthiness on perceivers' evaluations of targets, thus, findings were equivocal regarding the primary hypothesis. It is the thesis of this project that 1) highlighting experiences of the perceiver and the target, and 2) utilizing methods from developmental and experimental psychology, are both necessary to understand the broader implications of person perception research.

Implications of Person Perception Across Development: The Reciprocal Influences of Problem Behavior and Facial Trustworthiness

by
Zoe M. Alley

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented November 18, 2020
Commencement June 2021

Doctor of Philosophy dissertation of Zoe M. Alley presented on November 18, 2020

APPROVED:

Major Professor, representing Psychology

Director of the School of Psychological Science

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Zoe M. Alley, Author

ACKNOWLEDGEMENTS

The author expresses sincere appreciation to all parties that contributed, directly or indirectly, to this dissertation. Thank you to the scientists and staff at the Oregon Social Learning Center, for overseeing data collection of the longitudinal dataset that comprises Study 1 of the current project, for support in accessing those data, and for having the foresight to take pictures of participants in the Oregon Youth Study. Thank you to collaborators John Paul Wilson and Nick Rule for your assistance in processing and rating those photographs. Thank you to the lab assistants in the Youth Adjustment Lab and, in particular, thank you to Bryn Landrus for testing iterations of survey development and for your enthusiasm and passion for research. Thank you to the committee of this project, for your contributions to my development as a scholar, and for always making time to talk with me about science. Thank you to my advisor, David C. R. Kerr, who generously took in an "orphaned" graduate student upon returning from sabbatical, and who became a mentor whose guidance and friendship I will always cherish. Thank you to Bridget E. Hatfield, the brilliant scientist whose intellectual library of research made me realize that grad school was the path for me. I have striven to be the kind of scientist and the kind of person that you are. Thank you to the graduate students of the School of Psychological Science, whose advice, support, and friendship made this journey joyful. Some departments are marked by competition; ours has always been a place of friendship, of commiseration, of shared delight for science, of intellectual pontification and friendly debate. Our office, outfit with snack-drawer, microwave, and beanbag, was a second home to me. I am so proud to have you as colleagues, and even prouder to call you friends. Thank you to my family, my brothers, my sister, my parents, and my husband. Thank you for bringing me tea when I needed to study, or board games when I really needed a break. Thank you for being proud of me before I started this journey, and for staying proud of me every step of the way. Thank you for supporting me as a scholar, and for occasionally reminding me that I am a sister, a daughter, and a wife, too.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

DEDICATION

This dissertation is dedicated to my grandmother, Karen Carroll Fratzke. You showed up to celebrate my every accomplishment, with flowers. You were the bravest person I've ever known. I like to think it's your determination that gave me the strength and ambition to achieve this milestone.

Nana, this one's for you.

Implications of Person Perception Across Development: The Reciprocal Influences of Problem Behavior and Facial Trustworthiness

General Introduction

Adolescence and emerging adulthood are developmental periods characterized by heightened engagement in risk behaviors, such as delinquency and substance use (Jessor, 1991; Schulenberg & Maggs, 2002). During adolescence, these risk behaviors are powerfully influenced by environmental factors (Dick et al., 2016). Of particular importance for youth is the social environment; involvement with deviant peers, for example, is a strong predictor of substance use and delinquency (Dishion & Owen, 2002). A clear understanding of adolescents' social environment, including how important others respond to them throughout development, is essential to describe the development of problem behaviors and intervene with at-risk youth. Although traditionally siloed from health and developmental fields, research in person perception has compiled important theoretical models regarding how people perceive one another, and how such perceptions may influence the way perceivers behave toward and provide consequences for those around them (e.g., Ambady, Bernieri, & Richeson, 2000; Zebrowitz & Montepare, 2006). The application of this work to the study of developmental psychopathology has the potential to enrich both fields.

Despite the common adage "don't judge a book by its cover," people of all ages around the world make snap judgments about others' personalities based on facial appearance (Q. Li et al., 2017; Oosterhof & Todorov, 2008; Wilson & Rule, 2017; Zebrowitz & Montepare, 2008). Of particular social consequence is the evaluation of whether another person is benevolent and safe versus malevolent and dangerous—that is, how trustworthy someone appears (Oosterhof & Todorov, 2008). People with an untrustworthy facial appearance are treated more harshly than those who appear

trustworthy (Todorov et al., 2015). Some evidence indicates that facial cues regarding trustworthiness are at least partially associated with the behavioral history of the person being perceived: for example, from facial appearance alone, perceivers are able to distinguish between violent and nonviolent offenders with greater than chance accuracy (Stillman et al., 2010). There is also evidence that those who engaged in more delinquency and substance use during adolescence looked less trustworthy than their peers by early adulthood, according to strangers' ratings of their faces (Alley et al., 2019). Yet facial cues of trustworthiness are imperfect, and their influence on person perception can have disastrous consequences. For example, in a study of men who were falsely accused of murder and later exonerated, those who appeared untrustworthy were more likely to have received the death sentence (Wilson & Rule, 2015, 2016).

Thus, research in the field of person perception indicates that facial trustworthiness exerts social influence with potentially disastrous consequences for the person being perceived. Yet it is not known how a trustworthy appearance develops, why it is sometimes associated with antisocial behaviors, or how facial trustworthiness during the formative years of adolescence relates to long-term life outcomes in adulthood. To answer these questions, repeated measures of facial trustworthiness and relevant behavioral data on the targets of person perception across a wide developmental window are necessary. Although rare in the field of person perception (but see Q. Li et al., 2017; Zebrowitz et al., 1993), these longitudinal methodological practices are common in developmental psychology (Marcoulides, 2018). Thus, the integration of methodological, analytical, and theoretical tools from both developmental and social psychology are integral to the investigation of the reciprocal influences of appearance and

delinquent behavior, and the consequences of this relationship across development. This investigation would provide theoretical contributions to person perception and have practical implications for prevention and intervention from a developmental perspective.

Consequences of Appearance-Based Judgments for the Perceiver's Behavior

Person perception research concerns the processes by which individuals perceive and categorize one another (Ambady et al., 2000; Zebrowitz & Montepare, 2006). In terms of nomenclature, the person making a judgment about another person is labeled the "perceiver," and the person being perceived is the "target." Surprisingly, much of these person perception processes take place within the first few seconds after the perceiver sees a target, and these perceptions lead to consistent impressions between different perceivers and endure over time (Ambady et al., 2000). For example, student evaluations of teaching at the end of the term can be predicted from third parties' ratings of the teacher's nonverbal behavior, after viewing only seconds of footage from a lecture (Ambady & Rosenthal, 1993). Thus, social psychologists can study person perception by sampling "thin slices" of nonverbal social cues, such as a 10 second audio clip retaining only another person's vocal tone, a video of two people interacting with audio removed, or at the more extreme end, a still image of another person's face.

Since little time is required to form an impression that is reliable across perceivers and predictive of their future attitudes about the target, perceivers' impressions based on thin slices can be collected efficiently in a laboratory, and there is good reason to believe that those impressions will be generalizable to the impressions of many other people. For example, lab work regarding interpersonal judgments based on facial features indicates that those who appear childlike or "babyfaced" are judged as kind, honest, and naïve

(Berry & McArthur, 1985), those who appear attractive are perceived as more sociable and intelligent (Eagly et al., 1991), and, as mentioned before, those who appear trustworthy—as judged from the face—are perceived to be more worthy of trust than those who do not (Todorov et al., 2015; Wilson & Rule, 2017).

These rapid judgments (for trustworthiness, down to milliseconds; Todorov, Pakrashi, & Oosterhof, 2009) affect the way people treat one another. For example, in a lab task, men spoke more warmly over the phone to women whom they were led to believe were attractive (regardless of the women's actual appearance; Snyder, Tanke, & Berscheid, 1977). In an example from real court cases, babyfaced plaintiffs were less likely to be convicted of a crime involving malevolence than those with mature faces (Zebrowitz & McDonald, 1991). As another example, participants in a lab environment were less likely to invest money with someone who looked untrustworthy (van 't Wout & Sanfey, 2008).

Although these judgments are impactful, their accuracy is generally poor (Todorov et al., 2015), and sometimes entirely incongruent with the individuals' actual behavioral tendencies. For example, contrary to stereotypical expectations, babyfaced male adolescents performed better scholastically and behaved more aggressively than their mature-faced peers (Zebrowitz et al., 1998). Nevertheless, perceivers' prejudgments persevere (Wilson & Rule, 2017). In fact, perceivers struggle to attenuate these perceptions in favor of other information that may be more predictive of the target's future behavior, such as the target's previous behavior or environmental circumstances. For example, participants' investment decisions in a laboratory money-lending game, in which the target had the opportunity to exploit or reward the perceivers' trust, were

influenced more by photographs of the targets' faces than by whether the participant knew the target had an incentive to act selfishly (Jaeger et al., 2019). This was true even though participants reported that they believed those incentives were more diagnostic of prosocial behavior than facial cues. In the same paradigm, targets' previous history of trustworthy behavior only affected participants' behavior if this behavioral information preceded the presentation of the targets' face—that is, perceivers used facial cues to update their predictions for how the target would behave, but after seeing a face, anecdotal information about the target had no effect on their judgments (T. Li et al., 2017). However, perceivers were found to use behavioral information to update their impressions of the target's likeability, trustworthiness, and other general social evaluations after viewing a face if the target's reported behavior was extreme and diagnostic (e.g., saving a baby from a runaway train, mutilating a defenseless animal; Shen et al., 2020). Thus, judgments based on facial features are resistant to adjustment, but can be malleable in the face of convincing counter evidence.

Causes of Appearance-Based Judgments

Given the questionable accuracy of appearance-based judgments (e.g., Todorov et al., 2015), it is worth asking why they are pervasive, consistent, and socially influential. Zebrowitz and Montepare (2008) argue that appearance-based misjudgment occurs when an adaptive tendency is overgeneralized to inappropriate contexts: namely, using facial features (e.g., cues to age and approachability) to identify targets' affordances (Zebrowitz & Montepare, 2008). According to this overgeneralization hypothesis, babyfaced targets are perceived as naïve and benevolent because the configuration of their facial features triggers a readiness in perceivers to respond to actual babies, who are indeed naïve and

benevolent. Whereas perceivers' beliefs about targets on the basis of perceptions of babyfacedness is explained by an age overgeneralization, an emotion overgeneralization appears to underlie trustworthiness judgments. That is, participants judge facial features at a resting state that resemble smiling as more trustworthy and approachable, compared to facial features that at rest resemble a scowl (Oosterhof & Todorov, 2008).

Generally, it is appropriate to identify babies by their facial features and to treat those who are scowling with caution. Sometimes, those adaptive tendencies are overgeneralized to inappropriate contexts. Yet, the immediate drawbacks of a false positive for the perceiver (e.g., avoiding a benevolent stranger because he "looks shady") are less consequential than those of a false negative (e.g., approaching someone whose facial expression signals hostile intent; Zebrowitz & Montepare, 2008). Thus, such misperceptions persist.

Linking Targets' Appearance and Behavior

It is possible that an association between appearance and behavior could develop over time, such that facial characteristics may indeed be indicative of the person's behavioral tendencies. Zebrowitz' (1997) theoretical model outlines four potential mechanisms for such a link (see Figure 1). Briefly, an innate characteristic (e.g., elevated testosterone during adolescence; Pathway A) or environmental factors (e.g., poverty; Pathway B) may be a common cause of both appearance and behavior. The processes of primary interest for the proposed project involve the two other pathways, which emphasize the possibility of a direct causal link between behavior and appearance. Zebrowitz argues that facial appearance may constrain behavior through expectancy

effects (Pathway d-D), or that behavior might itself alter facial appearance (a Dorian Gray effect; Pathway C).

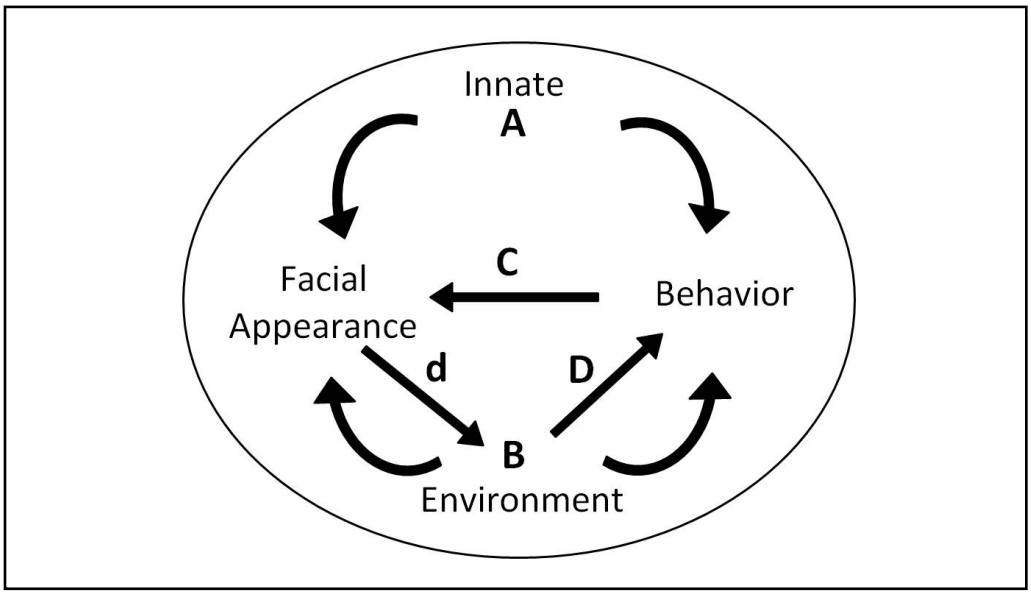


Figure 1. Model linking appearance and behavior, adapted from Zebrowitz, 1997.

Expectancy Effects (Pathway d-D)

Expectancy effects operate when the expectations of the perceiver elicit from the target the very behaviors the perceiver expected. The most famous example of these processes is the Pygmalion effect, in which a randomly selected subset of students outperformed their peers at the end of an academic year, only after their instructors were led to believe those students had the potential to excel (Rosenthal & Jacobson, 1968; see also Friedrich, et al., 2015; Raudenbush, 1984).

There is good reason to believe such effects could emerge from appearance-based judgments. For example, a target's likeability as judged from a photo predicted how warmly participants approached the target one month later, and also how much they actually liked the target upon meeting (Gunaydin et al., 2017). When speaking over audio only, men were warmer to women they were led to believe were more attractive,

and those women were rated by third party observers as more warm and likeable when their conversation partner believed they were attractive—even when the partner's audio was removed (Snyder et al., 1977). In this way, positive expectancies based on appearance elicited a positive behavioral response. It is possible that through many repeated, similar experiences, others' expectancies may be internalized by the target, such that the expected behavior occurs even when unprovoked by perceivers (see Figure 2).

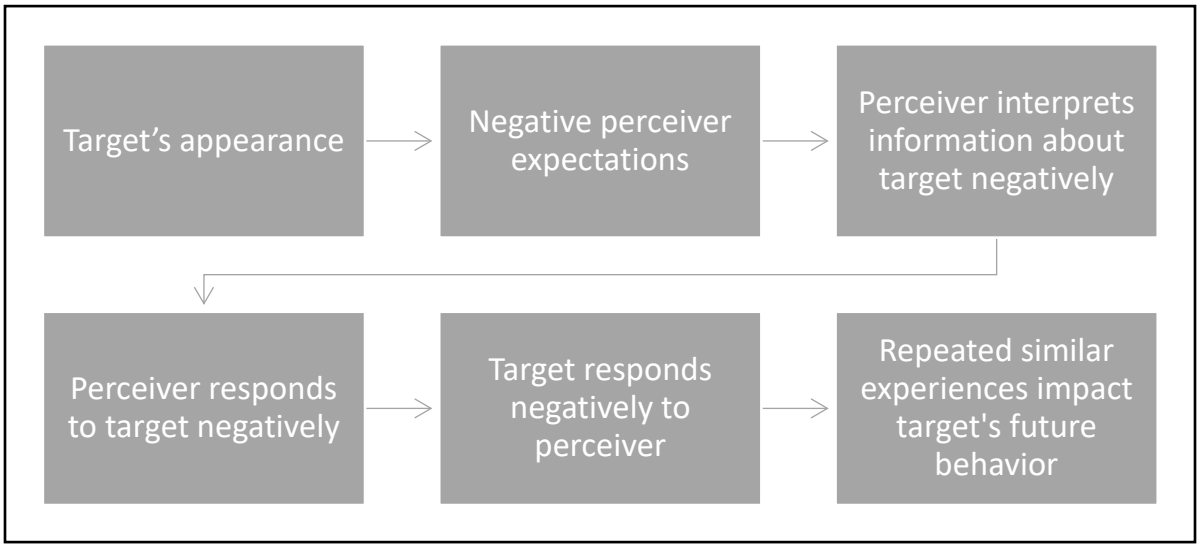


Figure 2. Model of appearance-based expectancy effects.

Expectancy effects may emerge in relation to facial trustworthiness, as those who appear untrustworthy elicit and experience a more a negative, suspicious, and limiting environment than those who do not. For example, perceivers' decisions are less influenced by a positive recommendation accompanied by an untrustworthy (versus a trustworthy) appearing face (T. Li et al., 2017). Furthermore, people respond more harshly to those who appear untrustworthy; in a laboratory task that allowed participants to deny themselves a reward in order to punish a partner who treated them unfairly, participants were more likely to punish targets who looked less trustworthy (Wu et al., 2018).

Thus, behavioral expectations based on facial trustworthiness affect perceivers' behavior toward the target in the short term. Preliminary evidence suggests that facial trustworthiness may predict targets' behavior in the long term. In a sample of school children ages 8-12 years, facial trustworthiness predicted peer acceptance, which in turn predicted students' behavioral trustworthiness one year later (as measured by peer report, which itself may be susceptible to facial trustworthiness; Q. Li et al., 2017). Thus, a person who appears untrustworthy may be treated with cautious distrust, and eventually may become a person worthy of suspicion.

The potential for expectancy effects based on facial trustworthiness is particularly troublesome when we consider adolescents. When perceivers have little behavioral information to draw on, such as when a new teacher is assigned or students begin to make friends in a new peer group, the potential for expectancy effects based on appearance may be particularly salient. As mentioned previously, adolescents are particularly susceptible to influence from their social relationships, especially when it comes to problematic behaviors such as substance use and delinquency. Expectancy effects regarding problem behaviors—which are relevant to perceptions of trustworthiness—may be some of the most dangerous and consequential, for both the person affected by expectancy effects and those affected by their problem behavior.

Measuring Trustworthiness Expectancy Effects. To determine whether appearance-based expectancy effects contribute to the development of delinquency, each link in the model portrayed in Figure 2 must be tested. First, it must be established that facial trustworthiness results in negative expectancies; as summarized above, this appears to be true. Further research should explore whether perceivers make different attributions

regarding those who appear untrustworthy (e.g., an untrustworthy appearing person who lies to avoid an uncomfortable situation may be perceived as dishonest, whereas a trustworthy appearing person who executes the same behavior may be perceived as tactful). Once this is established, a short-term interaction between perceivers and targets could be assessed in a laboratory environment, where perceivers' expectancies are manipulated (as in Snyder et al.'s study, 1977, wherein perceivers' impression of the targets' attractiveness predicted the targets' warmth as reported by third party observers). Such a design would determine how perceivers' perceptions of the targets' trustworthiness alter their behavior toward the target, and how this influences the targets' concurrent behavior. None of these methodologies demonstrate whether numerous similar encounters may impact targets' behavior and outcomes in the long term.

A panel design that assesses appearance, important others' first impressions based on facial appearance (e.g., by having teachers rate incoming students' photographs before beginning a new school year), and problem behaviors could examine the potential for enduring expectancy effects by analyzing change in behavior across time. In contrast, a longitudinal design that follows participants for many years would allow for an investigation of 1) the development of facial trustworthiness across time and 2) the development and stability of behavior congruent with expectancies related to an untrustworthy appearance.

The present study investigated expectancy effects by assessing the association between early facial trustworthiness and engagement in delinquent behavior across time through a longitudinal, developmental lens. It should be noted that expectancy effects may be affected by numerous, brief interpersonal encounters throughout day to day life

that would not be captured in such a design. To better understand some of the intermediary processes between target's appearance and perceiver's response, this longitudinal analysis was paired with a laboratory study that examines perceivers' reactions to ambiguous behavioral information about those who vary in facial trustworthiness (the third box in the model depicted in Figure 2).

The Dorian Gray Effect (Pathway C)

Engagement in stereotypically untrustworthy behavior may have a deleterious effect on one's appearance. For example, someone who is aggressive may make angry facial expressions often, which may cause wrinkling or muscular changes that result in a face that appears unapproachable even at rest. Indeed, there is some evidence that perceivers' personality judgments of older adults' faces, who have had the most time for habitual facial expressions to impact their facial appearance, are associated with self-reports and romantic partners' reports of actual personality (Malatesta et al., 1987).

Antisocial behaviors such as theft, vandalism, and violence are of great relevance to perceptions of trustworthiness. It is possible that aggressive facial expressions or engaging in antisocial behaviors may lead directly to an untrustworthy appearance (e.g., a broken nose from a physical altercation). Yet, particularly during the developmental period of adolescence, delinquency tends to covary with separate but related behaviors (e.g., substance use; Jessor, 1991), which may also influence facial appearance through independent channels (e.g., Okada et al., 2013). Alley and colleagues (2019) found evidence that engagement in delinquent behaviors during adolescence was associated with facial trustworthiness at adulthood, after controlling for facial trustworthiness at early adolescence. However, this relationship was primarily explained by elevated

tobacco use, which was more common among those who reported delinquent behaviors. Thus, a Dorian Gray effect may emerge when those who engage in antisocial behaviors use more tobacco, which results in deleterious effects on facial trustworthiness. Such a pathway would also imply that individuals who use tobacco but do not engage in those behaviors would be left with an undeservedly untrustworthy appearance.

Adolescence is a period of heightened engagement in risk behavior (Jessor, 1991). If such behaviors have deleterious effects on facial appearance, youth who temporarily engaged in substance use or crime during adolescence may continue to be socially disadvantaged even after problem behaviors have ceased. Therefore, it is important to determine the extent to which engagement in problem behaviors affect appearance during adolescence, as an untrustworthy appearance may lead to social disadvantages (e.g., reduced employment opportunities) later in life.

Measuring the Dorian Gray Effect. A laboratory environment is ill suited for research on the Dorian Gray effect. Malatesta and colleagues (1987), for example, found that perceivers' impressions of targets' personalities, based only on a facial photograph, were associated with targets' self-reported personalities; although this study demonstrates that a relationship between appearance and personality existed in this sample, it cannot explain how that relationship developed over time. Appearance data from at least two time points are required in order to demonstrate that appearance has in fact changed over time, and thus distinguish Dorian Gray effects from other causes (e.g., Pathways A or B). Sampling a multitude of developmental periods allows for the definition of the shape of change in appearance across time, and how this may relate to targets' behaviors. Although longitudinal methodology does not rule out the possibility that some third

variable elicited both untrustworthy behavior and changes in appearance across time, evidence of an association between such changes and delinquency would be strong evidence for a Dorian Gray effect. Thus, the present study paired photographs across adolescence and into adulthood with prospectively assessed report of delinquent behaviors, in order to investigate how changes in appearance across time might relate to behavior.

The Present Project

Both Dorian Gray and expectancy effects imply change across time (in either behavior or appearance). Yet there is little research that extends the implications of facial trustworthiness to participants' lives by applying this model to a longitudinal, developmental context. Methodological tools from developmental psychology, which allow for the simultaneous analysis of development in both appearance and behavior, are particularly appropriate to fill this gap.

The goals of the present project were to evaluate the stability of facial trustworthiness from adolescence to early adulthood, to determine the reciprocal associations between an untrustworthy appearance and related behavior, and to explore temporal associations consistent with Dorian Gray and expectancy effects. To address these goals, two studies utilizing two different samples and methodologies were employed.

Study 1 focused on the targets of perceptual judgments. The development of an untrustworthy facial appearance and its relationship to delinquency was examined in the participants of the Oregon Youth Study (OYS), a preexisting dataset that followed approximately 200 males selected from neighborhoods with the highest local rates of

police reported juvenile delinquency, from ages 10 to age 38 (Capaldi & Patterson, 1989). This dataset captures the critical developmental periods of adolescence and early adulthood in an at-risk sample, such that a range of delinquent behaviors were represented. This sample permits the specification of the development of facial trustworthiness from ages 13 to 38, the examination of the extent to which initial levels of facial trustworthiness predict rate of change in self-report of delinquent behavior, and furthermore, the extent to which delinquency predicts rate of change in facial trustworthiness. Few if any prior studies have investigated simultaneously the development of both risk behaviors and facial appearance.

Study 2 focused on the influence of targets' facial trustworthiness on perceivers' interpretation of ambiguous behavioral information about the target, as bias in this process may lead to expectancy effects. The purpose was to elucidate a preliminary process that may link the targets' facial trustworthiness with later outcomes. In this lab study, participants viewed faces selected from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015) that varied in facial trustworthiness. Faces were paired with a brief vignette about the target engaging in ambiguous behaviors that were either clearly assertive and could be interpreted as hostile (based on the Donald vignette; Srull & Wyer, 1981) or clearly passive with no allusions to hostility. Then, participants reported on their assessment of the perceived hostility, trustworthiness, and likeability of the target, and on how likely they would be to engage with the target themselves. Faces that appeared less trustworthy were hypothesized to be rated as less friendly, kind, considerate, thoughtful, likeable, and trustworthy. Critically, the effect of facial trustworthiness on perceivers impressions of the target's personality was predicted to be

stronger when the target's face was paired with the assertive, ambiguously hostile vignette than when the target's face was accompanied by the passive vignette.

Study 1

Facial Trustworthiness Predicts Rate of Change in Delinquency Among At-Risk Men from Ages 13 to 38 Years

The social environment has a striking influence on peoples' engagement in delinquency, particularly during adolescence (Dishion et al., 1999; Jessor, 1991). In order to reduce and prevent delinquency, which has significant individual and societal costs, its social antecedents and consequents must be understood.

Nonverbal cues such as vocal tone, body language, and appearance can greatly influence the social environment (e.g., Ambady et al., 2000; Ambady & Rosenthal, 1993; Snyder et al., 1977), yet there has been a dearth of research on the relationship between such nonverbal cues and delinquency. Facial cues regarding emotional expression (e.g., smiling versus scowling) may signal whether a potential social partner is benevolent or threatening. Although leveraging such facial cues to guide behavior is generally adaptive, this tendency can be overgeneralized to inappropriate contexts. Since downturned lips are characteristic of a scowl, an individual with naturally downturned lips at rest may be judged as hostile by potential social partners when making no facial expression at all. This phenomenon is termed the emotion overgeneralization effect (Zebrowitz & Montepare, 2006). Emotion overgeneralization appears to contribute to perceivers' assessment of another person's trustworthiness, with those who have faces resembling a smile judged as more trustworthy than those who have faces that at rest resemble a scowl (Todorov et al., 2008).

The person perception literature contains ample evidence that facial trustworthiness of the target is associated with perceivers' assessments of their personality (Todorov et al., 2015; Wilson & Rule, 2017) and perceivers' behavior toward

the target in lab tasks (T. Li et al., 2017). *In vivo,* facial trustworthiness has been associated with peer relationships and popularity in grade school (Q. Li et al., 2017), and sentencing outcomes in adulthood, such that those who appeared less trustworthy were more likely to receive the death sentence—even among those who were later exonerated (Wilson & Rule, 2015, 2016). Just as teachers' expectations regarding their students' aptitude to learn influenced their students' actual learning in Rosenthal's famous study (Rosenthal, 1987; Rosenthal & Jacobson, 1968), it may be that perceivers' expectancies originating from a particularly trustworthy or untrustworthy appearing person elicit from the target confirmatory behaviors. Because how trustworthy a person appears relates to the perceiver's perception of the target's capacity to take advantage of or harm others, these expectancies relate to perceptions of the target's propensity to engage in delinquent behaviors, such as theft, vandalism, and violence (Flowe, 2012).

It is possible, then, that those who appear untrustworthy are perceived by others as more likely to become delinquent, and therefore their appearance may elicit an environment that facilitates engagement in delinquent behaviors. In this way, facial appearance (in particular, facial trustworthiness) may exacerbate or attenuate the development of delinquency across time. Social expectancies related to appearance may be particularly influential for delinquent behavior that occurs during adolescence, when social factors are a strong predictor of risk-behavior and delinquency (Dishion, 2000; Dishion & Owen, 2002). As delinquent behaviors are associated with a host of negative outcomes for both the actor and those negatively impacted by those delinquent acts, identifying and understanding potential influences on the development of delinquency across time is paramount.

Adding further complexity to the issue, adolescents' delinquent behaviors may impact facial trustworthiness directly (e.g., through increasing opportunities for negative facial expressions that accumulate in changes in the wrinkling and musculature of the face), or indirectly, through other variables that covary with delinquency. In a longitudinal study among at-risk boys, delinquency was associated with decrements in facial trustworthiness from ages 14 to 24 years (Alley et al., 2019). This association may have been mediated by tobacco use, which covaried with delinquency, and has been shown to affect wrinkling around the eyes and lips that may influence perceived facial trustworthiness (Okada et al., 2013). Decrements in facial trustworthiness resulting from an individual's delinquency or substance use accumulated across adolescence may lead to social disadvantages as an adult, even among those who have desisted in such behaviors. Furthermore, the social consequences related to an untrustworthy appearance may themselves serve as barriers to desistence, e.g., by decreasing opportunities to become gainfully employed or to develop a prosocial friend group.

The present study leveraged photographs from an at-risk sample of boys followed from ages 13 to 38 years for two purposes: first, to explore the nature of change in facial trustworthiness across development (which is as of yet undefined); second, to relate the developmental trajectory of facial trustworthiness to delinquency across this window, in order to test for Dorian Gray (behavior shaping the appearance) and expectancy effects (appearance shaping behavior).

Hypotheses

H1. Describe patterns of change in facial trustworthiness across adolescence and adulthood.

Development in facial trustworthiness from ages 13 to 38 years was explored. As no prior work had investigated the development of facial trustworthiness across such a wide developmental period, it was unclear whether trustworthiness would decrease, increase, or remain stable on average across time. However, babyfacedness, a trait associated with trustworthiness, decreased across time among men (Zebrowitz et al., 1993). In addition, Alley and colleagues (2019) found evidence that trustworthiness decreased across ages 14 to 24 in this sample, at least among men who engaged in more tobacco use and delinquency than their peers. Furthermore, given that perceptions of trustworthiness involve an assessment of how threatening versus approachable the target may be, it is also likely that those who appear younger may be perceived as less threatening and therefore more trustworthy. Finally, given the rapid physical development characteristic of adolescence (Marečková et al., 2011), changes in facial trustworthiness may be more dramatic during adolescence than adulthood.

H1.1 (Exploratory). Perceived facial trustworthiness is expected to decline across ages 13 to 38, such that participants are perceived as progressively less trustworthy as they age.

H1.2 (Exploratory). Declines in facial trustworthiness will become less pronounced or stabilize across adulthood.

H1.3 (Exploratory). There will be significant variance in intercept and slope factors, such that there is substantial individual variation in rate of change and initial levels of facial trustworthiness.

H2. Determine the relationship between delinquency and facial trustworthiness across development.

Evidence of the Dorian Gray effect, in which an individual's behavior alters their facial appearance over time, and expectancy effects, in which the target's facial appearance creates a social environment that elicits behavior consistent with their appearance, was assessed. These effects are not mutually exclusive. Note that H2 involves the assessment of targets' facial features and targets' behaviors, but not perceivers themselves. These crucial mechanisms thought to underlie expectancy effects were investigated in Study 2.

Evidence of expectancy effects and/or the Dorian Gray effect may reflect social pressures that increase adolescents' delinquency, or consequences of risk behaviors during a critical developmental period that could negatively impact a person's life even after those behaviors have ceased. The extent to which the patterns of change in facial trustworthiness identified during exploratory analyses in H1 and initial levels of facial trustworthiness at age 13 years are associated with delinquency over time will be assessed, in order to investigate the phenomena described above.

H2.1. Initial levels of delinquency will predict rate of change in facial trustworthiness from ages 13 to 38, such that those with higher initial levels of delinquency show greater decreases in facial trustworthiness (Dorian Gray effect).

H2.2. Initial levels of facial trustworthiness will predict rate of change in delinquency from ages 13 to age 38. That is, those who initially appeared less trustworthy will increase in delinquency more rapidly or will decline in delinquency more slowly than those who appear more trustworthy (expectancy effects).

Method

Participants

Participants were drawn from the Oregon Youth Study, a longitudinal study of at-risk boys that started in the mid-1980s (see Capaldi & Patterson, 1989). Boys were recruited in entire fourth-grade classrooms from schools in neighborhoods with the highest rates of police-reported delinquent episodes by juveniles. Parents of 74% of the targeted boys allowed their son to participate (Capaldi & Patterson, 1989). Families, and later the participants themselves, were provided $100 for participating in annual interviews (Capaldi et al., 1997). The Oregon State University Institutional Review Board ceded oversight of this study to the Oregon Social Learning Center (OSLC), which conducted the Oregon Youth Study. Thus, the present study was approved for human subjects research by the OSLC Institutional Review Board.

The sample was 90% White (3% African American, 2% American Indian, 1% Mexican American, and 5% other identities) and largely from low socioeconomic families. Median annual income at study entry was $15,000. Multimethod, multiagent assessments of participants took place regularly from ages 10-38 years.

Measures

Control Variables

Smiling. Although participants were instructed to maintain a neutral expression during photographs, many of them smiled. Given the association between smiling and facial trustworthiness identified in prior studies (Krumhuber et al., 2007; Ozono et al., 2010), undergraduate psychology students were recruited to rate smiling for each wave of photographs. These ratings were used as a control. Smiling was measured on a 1 to 5 scale, where each value was linked to a discrete expression: 1 = no smile, 2 = slight smile, 3 = smile (no teeth showing), 4 = smile (teeth), 5 = full smile. Raters for each wave of photographs ranged from $k = 2 – 5$ raters. Where k was greater than 2, Cronbach's alpha was calculated as a measure of reliability. In these cases, alpha ranged from .93 to .97. In the two instances where $k = 2$, $r = .80$ - $.82$ ($p < .001$).

Parent income. Parents reported their annual family income during assessments at participants' ages 10, 11, and 12 years ($r = 0.79$ - 0.88). Income was z-scored at each timepoint and all three observations were averaged by study staff, producing a single score for each participant. These archival variables were utilized as a proxy for socioeconomic status in the following analyses.

Facial Trustworthiness

Study staff took photographs of the participants annually at ages 13-18, and then at 21, 24, 32, 36, and 38 years. Research assistants in collaborating labs at University of Toronto and Montclair State University cropped the photographs around the face and rendered them in black and white. Most photographs were cropped close around the face

and excluded hairstyle; however, photographs at age 24 years (the first wave prepared for analyses) were cropped in a square and included hairstyle.

Undergraduate psychology students blind to the study hypotheses and the basis for OYS recruitment coded photographs for perceived trustworthiness on a scale from 1 to 7 (1 = very untrustworthy, 7 = very trustworthy). Raters were sampled until ratings converged on a reliable criterion ($\alpha \geq .80$). Between 12 and 25 raters rated each face, and an average facial trustworthiness score was calculated for each photograph and served as the criterion variable.

Not all participants had photographs available at each wave; therefore, the number of participants with facial trustworthiness ratings varied over time. Of 206 participants, 30% ($N = 63$) were photographed at all 10 timepoints, 85% ($N = 195$) were photographed during at least 7 timepoints, 97% ($N = 201$) were photographed during at least 5 timepoints, and at least two photographs were available for all participants. At least two photographs during adolescence in addition to at least two photographs during adulthood (ages 18-38) were available for 197 participants, permitting growth to be estimated in both developmental epochs.

Coding of photographs was completed for all ages except age 32 years. Closure of the university in March 2020 due to the COVID-19 pandemic precluded recruitment of student raters on campus for this wave, and the sensitive nature of these photographs precluded moving such recruitment online. Raters for each wave of facial trustworthiness ranged from $k = 10$-25. Alpha for facial trustworthiness was at .79 or higher for all waves.

Delinquency

Participants completed the Elliott Delinquency Scale during annual assessments (Elliott et al., 1983) at ages 13 to 26 years and 28 through 32 years, and then at ages 34, 36, and 38. Designed as an analogue to the Federal Bureau of Investigation's Uniform Crime Reports arrest measure, the scale assesses frequencies with which participants engaged in a range of antisocial behaviors during the prior year (e.g., theft, vandalism, and violence). Example items include: [How many times in the last 12 months have you…] "…failed to return extra change that a cashier gave you by mistake," "…knowingly bought, sold or held stolen goods or tried to do any of these things", and "…attacked someone with the idea of seriously hurting or killing that person?" Responses for each item were capped at 365 to reduce skew and the influence of specific years on the total score. Prior work has established the reliability and validity of this measure (Elliott et al., 1983).

Data Analysis

To characterize initial levels and change across time in facial trustworthiness and delinquency over ages 13 to 38 years, a series of latent growth curve models were run using Mplus (Muthén & Muthén, 1998-2017)[1].

Indices of Model Fit

Model fit for exploratory hypotheses (H1.1-3) was evaluated via the χ^2 statistic, Comparative Fit Index (CFI; Bentler, 1990), Tucker Lewis Index (TLI; Tucker & Lewis, 1973), and root mean square error of approximation (RMSEA; Steiger, 1990). Although nonsignificant χ^2 values are indicative of good fit, this index is particularly sensitive to large sample sizes and is thus best interpreted in the context of other indices of model fit.

---

[1] Traditionally, person perception has utilized statistical methods from experimental psychology to investigate changes in appearance across time. For example, Zebrowitz and colleagues (1993) analyzed the stability of babyfacedness and attractiveness in a longitudinal sample that followed participants from ages 9 to 56 years of age through a series of ANCOVAs and t-tests. Although such methods permit the identification of significant overarching effects (e.g., perceptions of babyfacedness varied by age group) and the examination of whether appearance at each age was significantly different from appearance at the preceding and following age, they are not without limitations. Numerous statistical tests are required, any one anomalous wave of assessment may disguise subtle trends over time, and missingness can vary greatly across waves, biasing results (Graham, 2009). Longitudinal growth modeling, a statistical method common in developmental psychology, synthesizes a series of observations into latent overarching parameter estimates, i.e., an *intercept* and *slope.* The intercept represents the level of a given variable, taking into account variance from all timepoints of assessment (e.g., the intercept may represent latent average levels of facial trustworthiness at the first wave of observation). The slope represents rate of change across time, clocked from the intercept (i.e., to what extent facial trustworthiness increases/decreases across time). The complexity of the model can be adjusted in order to best fit the data. For example, the addition of a quadratic term can modify a linear slope estimate; a positive quadratic term in a model with a positive linear slope would indicate that the variable of interest increases across development, and that those increases escalate over time, such that change in appearance is most rapid and pronounced at later years. As another example, a piecewise model can be used to estimate two slope terms, such that a variable may increase across ages 13 to 18 and then decrease across the remaining ages.

Thus, a longitudinal growth model synthesizes multiple observations into significantly fewer statistical tests. Furthermore, it permits the use of maximum likelihood estimation where data are missing, allowing for stronger estimation than that permitted by listwise deletion (Graham, 2009). Finally, error is minimized across all observations, such that each timepoint contributes to an overall model representing growth across time. Given that the present study integrates numerous observations across a wide developmental window with many opportunities for missingness, a growth modeling framework is the most parsimonious method of interrelating trends across time in facial trustworthiness and delinquency.

Adequate fit is indicated by CFI ≥ .90, TLI ≥ .90, and RMSEA < .08, and good fit is indicated when CFI ≥ .95, TLI ≥ .95, and RMSEA < .05 (Schweizer, 2010). These benchmarks were treated as guidelines for assessing the extent to which models represent the data well, where models that approach the cutoffs were favored over models that did not.

Treatment of Missing Data

Maximum likelihood estimation, which has been shown to produce less biased estimates than listwise deletion, was utilized to account for missing data (Graham, 2009). Missing values for facial trustworthiness and delinquency were coded at -999. Missing values for smiling, a time varying covariate, were coded at 2, a plausible value on the 1 to 5 smiling scale (Muthen, 2008). Smiling variables were not given the missing value of -999 because, if treated as missing, Mplus used listwise deletion for all participants for whom any one smiling rating was missing. Smiling was missing for each instance where facial trustworthiness was missing, since smiling codes originated from the same photographs as facial trustworthiness ratings. Therefore, listwise deletion resulted in a sample of $N = 64$. Smile ratings were given a code of 2, which did not match a known missing value; thus, cases missing on smiling were not deleted. Because those values were ignored when the variable of interest (facial trustworthiness) was missing, the unique missing value for time-varying covariates did not contribute to model estimation and allowed for analyses to include the full sample of 206 participants.

Data Analysis Plan for Hypothesis 1

Exploratory hypotheses of the initial levels and change across time of facial trustworthiness (H1) were examined from ages 13 to 38 years. Initially, a visual

inspection of raw scores across time indicated potential appropriateness and feasibility of fitting a linear growth model. Next, an intercept only model where the slope term was constrained to zero was tested. Then, growth terms (linear, quadratic, etc.) were added incrementally until the model reached good fit, as evidenced by the indicators defined above.

As ages 13-38 included multiple developmental epochs (adolescence and adulthood), it was possible that rate of change in the variables of interest differed by epoch. For example, from ages 13 to 17 (adolescence), changes in facial trustworthiness may be more rapid and pronounced than across ages 18 to 38 (adulthood) due to physical maturation. Thus, a piecewise latent growth curve model was tested to parse growth in adolescence and adulthood. These models allowed for estimation of independent slope terms for these two developmental periods. The developmental "knot", or the point at which slope changes in the model to arrive at best fit, was hypothesized to be between ages 18 and 21.

There is little to no prior research investigating the development of facial trustworthiness. As exploratory hypotheses, facial trustworthiness was predicted to decrease across time (H1.1). Facial trustworthiness was expected to decrease less or stagnate during adulthood (H1.2). Finally, substantial variance was expected in the intercept and slope terms (H1.3), meaning that levels in early adolescence and change in facial trustworthiness across time were expected to vary between participants.

Data Analysis Plan for Hypothesis 2

Once a univariate model for facial trustworthiness was defined, a univariate latent growth curve model of delinquency was constructed. Rather than exploring the shape of

delinquency, which has been addressed by numerous prior researchers (e.g., Chen, 2010; Patterson, 1993; Patterson et al., 1989; Wiesner & Capaldi, 2003), the model formulation goal for delinquency was to develop a parsimonious model that 1) fit the data adequately and 2) would coordinate well with the final facial trustworthiness growth model. For example, if using a piecewise model, matching the developmental "knot" (or the timepoint when slopes change) across delinquency and facial trustworthiness would allow the latent slopes to refer to change across the same period of time in both models, facilitating more meaningful interpretations of model output.

A parallel process growth model was utilized to examine how initial levels and rate of change in delinquency and trustworthiness interrelated across ages 13 to 38. Specifically, age 13 levels of delinquency (intercept) were predicted to be negatively associated with slope of facial trustworthiness (change across time; H2.1, consistent with Dorian Gray effects), and age 13 levels of facial trustworthiness (intercept) were expected to negatively predict slope in delinquency (H2.2, consistent with expectancy effects).

Note that a negative association between intercept and slope implies that at higher levels of the intercept, the slope is *more negative.* The specific interpretation of such a trend depends on the value of the slope. If the slope is negative, individuals with higher levels of the variable of interest show greater decreases across time than those with lower levels. If the slope is positive, individuals at higher levels of the intercept increase more slowly than their peers at lower levels of the intercept.

Results

Descriptive Statistics

Control Variables

Smiling. Smile ratings were collected at all waves that were rated for facial trustworthiness. Therefore, patterns of missingness for smiling are the same as for facial trustworthiness (see below). Smile ratings ranged from 1 (no smile) to 5 (full smile) across all waves (except age 36, where smile ratings did not exceed 4.6). Mean smile ratings ranged from 1.66 to 2.16 across all waves. Smiling was not consistent within person across waves (e.g., some boys who smiled during age 13 photographs did not necessarily smile in age 18 photographs), suggesting that smiling should be controlled when specifying within-person change in facial trustworthiness.

In growth models, where smiling scores were time-varying covariates of facial trustworthiness, smiling scores were centered at 1 such that the parameter estimates could be interpreted at a smiling level of 1 (i.e., *no smile*), rather than 0, a score that has no meaning on the 1-5 smiling scale utilized in the present study.

Parent income. Parent income ranged from -2.12 to 2.12, and as would be expected given that the variable was z-transformed, the variable had a mean of 0 and *SD* of 0.9.

Facial Trustworthiness

Facial trustworthiness scores for participants ranged from 1.58 to a maximum of 5.59 across all waves. It may appear that raters were not using the full rating scale, which ranged from 1 (very untrustworthy) to 7 (very trustworthy), where 4 is neutral. However, the facial trustworthiness score for each participant is an average of *k* raters' ratings, such

that individual participants' scores are pulled away from the ends of the scale. Individual raters did indeed utilize the extreme ends of the scale. Between waves, mean facial trustworthiness ratings ranged from 3.22 to 3.48, just below the midpoint of the scale. Although there was variation across time in total facial trustworthiness score, the standard errors across time remained nearly equal (*SE* = .04 - .06; see Table 1).

Table 1. *Descriptive statistics and reliabilities for facial trustworthiness*.

Age	*N*	Mean	*SE*	Min	Max	*k*	Alpha
13	183	3.84	0.05	2.00	5.53	19	.83
14	175	3.79	0.05	2.20	5.25	20	.86
15	144	3.83	0.06	2.29	5.59	17	.87
16	173	3.38	0.05	2.00	5.00	15	.80
17	190	3.52	0.05	1.86	5.19	16	.84
18	179	3.49	0.05	1.88	5.13	21	.87
21	161	3.11	0.05	1.62	4.92	13	.83
24	177	3.76	0.04	2.40	5.20	19	.85
36	173	3.22	0.05	1.58	5.08	12	.81
38	169	3.43	0.04	1.95	5.00	18	.79
N (listwise)	63						

Note: *k* = Number of raters whose ratings were averaged to construct the facial trustworthiness score for each participant in that wave. Alpha represents internal consistency across raters.

Facial trustworthiness scores appeared to decline over adolescence (see Figure 3). The pattern of change in facial trustworthiness across adulthood was less clear. Age 24 facial trustworthiness was higher than the two preceding and following timepoints, disrupting an overall pattern of decline and then stabilization.

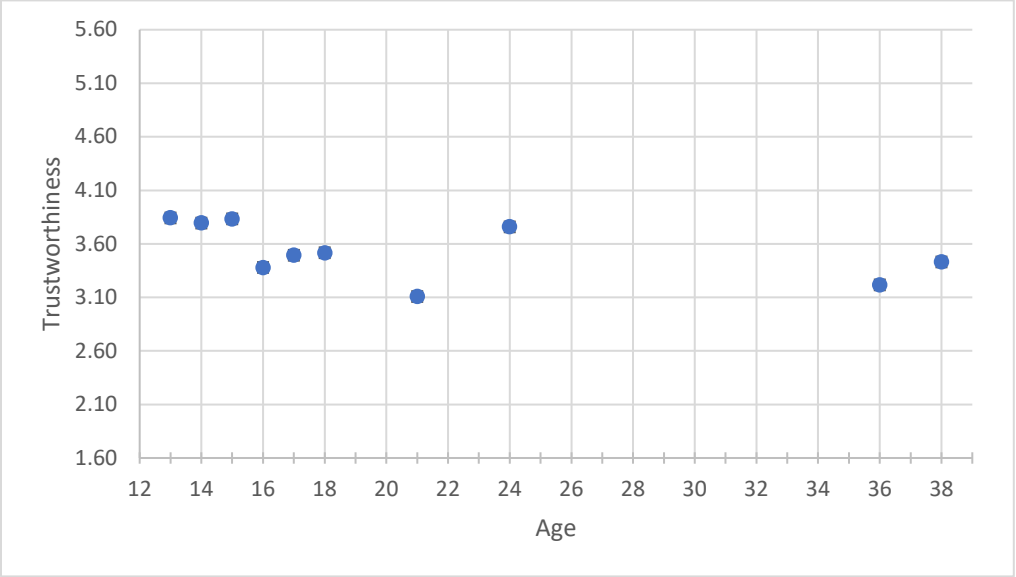


Figure 3. Facial trustworthiness across time, unadjusted for smiling. *N* ranged from 144 to 190. The Y axis was set to reflect the actual range of participants' scores (an average of *k* raters' ratings; see Table 1), rather than the full range used by individual raters (1 to 7).

Because smiling inflates perceptions of facial trustworthiness (Krumhuber et al., 2007; Ozono et al., 2010), and because levels of smiling were higher in earlier ages, change in raw facial trustworthiness scores across time was also examined excluding all instances where smiling met or exceeded 3 (smile, no teeth showing) on the 1-5 smiling scale (see Figure 4). Trends across time were comparable in this subset; notably, average facial trustworthiness at age 24 was higher than any other timepoint. This may be because the methods of rating age 24 facial trustworthiness were slightly different than other waves; i.e., age 24 photographs were cropped in a square shape and included hairstyle, whereas all other photographs were cropped in a circle around the face, excluding hairstyle.

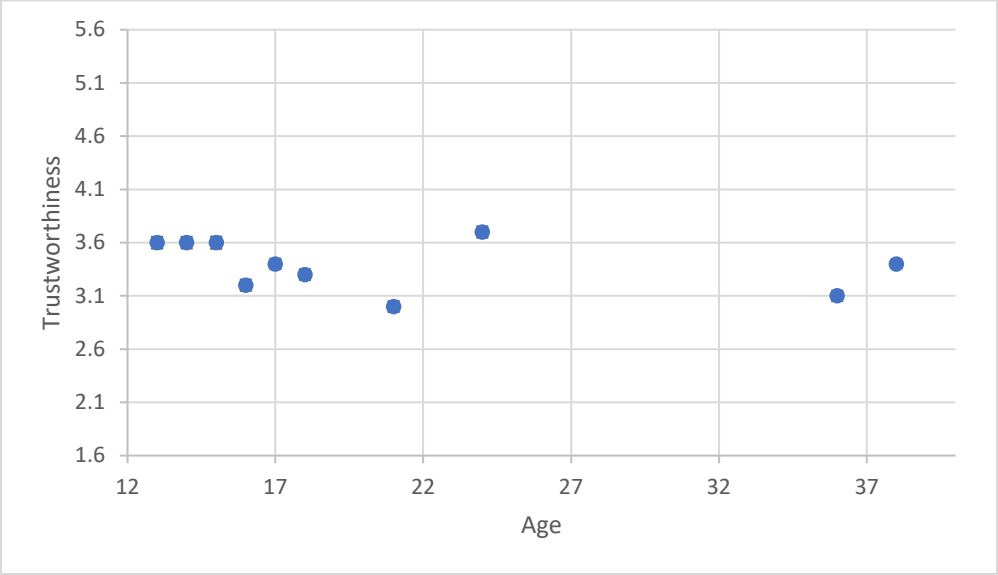


Figure 4. Exploratory graph of facial trustworthiness across time among cases with a smiling score less than 3. *N* ranged from 112 to 155. The Y axis was set to reflect the actual range of participants' scores (an average of *k* raters' ratings; see Table 1), rather than the full range used by individual raters (1 to 7).

Given the anomalous value and methodological differences in age 24 facial trustworthiness ratings, this wave was excluded from the following analyses. When age 24 was excluded, patterns of missingness were comparable with that of the complete dataset: 30% ($N = 63$) of participants had photographed at all 9 timepoints, 76% ($N = 157$) were photographed during at least 7 timepoints, 96% ($N = 198$) were photographed during at least 5 timepoints, and at least two photographs were available for all participants. At least two photographs during adolescence in addition to at least two photographs during adulthood were available for 95% of participants ($N = 194$), permitting growth to be estimated in both developmental epochs.

Delinquency

For self-reported delinquency, sample sizes by assessment wave ranged from 164 to 204. Complete data (22 assessments) were available for 62% of participants ($N =$

128), 20 observations or more were available for 85% of participants ($N = 177$), less than 20 waves of assessment were available for 20% of participants ($N = 29$), and only 11 participants had fifteen waves of assessment or fewer. At least two waves of assessment during both adolescence and adulthood were available for all but one participant, for whom no delinquency data were available during adolescence. The lowest assessment frequency for delinquency for any participant was 9 waves of assessment, which applied to only one participant.

Delinquency was not rare in this at-risk sample. At age 14, roughly 68% of the sample had reported at least one delinquent act, and at age 34, when delinquency was least common, 37% of the sample reported at least one delinquent act (see Figure 5).

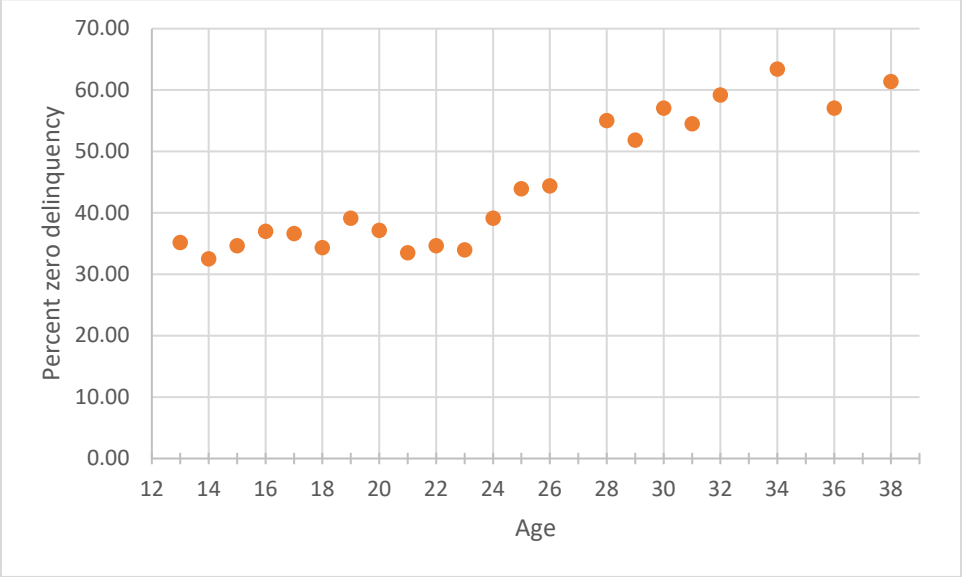


Figure 5. Percent of participants reporting no delinquent activities at each wave. *N* ranges from 164 to 203.

Mean scores for delinquency ranged from 0.61 ($SE = .09$) to 3.27 ($SE = .29$). Delinquency was right skewed, such that many participants reported zero delinquent activities at each wave, and the most extreme high scores were least common. Of note,

the maximum or most extreme delinquency score generally decreased with participants' age (see Table 2).

Table 2. *Descriptive statistics for delinquency.*

Age	*N*	Mean	*SE*	Min	Max	Freq 0	% 0
13	199	2.47	0.26	0.00	25.00	70.00	35.18
14	203	2.53	0.23	0.00	15.00	66.00	32.51
15	202	2.52	0.29	0.00	23.00	70.00	34.65
16	200	2.82	0.31	0.00	23.00	74.00	37.00
17	202	3.24	0.34	0.00	25.00	74.00	36.63
18	201	3.27	0.29	0.00	19.00	69.00	34.33
19	202	2.89	0.29	0.00	24.00	79.00	39.11
20	202	2.71	0.27	0.00	26.00	75.00	37.13
21	203	2.26	0.21	0.00	19.00	68.00	33.50
22	202	2.14	0.19	0.00	18.00	70.00	34.65
23	203	1.96	0.19	0.00	16.00	69.00	33.99
24	202	1.58	0.16	0.00	19.00	79.00	39.11
25	198	1.44	0.15	0.00	12.00	87.00	43.94
26	196	1.40	0.15	0.00	13.00	87.00	44.39
28	189	0.99	0.12	0.00	12.00	104.00	55.03
29	191	0.85	0.10	0.00	10.00	99.00	51.83
30	191	0.88	0.10	0.00	7.00	109.00	57.07
31	189	0.77	0.09	0.00	8.00	103.00	54.50
32	191	0.97	0.16	0.00	22.00	113.00	59.16
34	164	0.61	0.08	0.00	5.00	104.00	63.41
36	184	0.75	0.11	0.00	15.00	105.00	57.07
38	176	0.61	0.09	0.00	10.00	108.00	61.36
N (listwise)	128						

Note: Freq 0 = Number of participants who reported no delinquency during that wave. % 0 = Percent of participants within that wave who reported no delinquency.

On average, self-reported delinquency appeared to increase in adolescence and peak at age 18. Delinquency then decreased across ages 18 to 38, and those decreases decelerated with time, with the most rapid decreases in delinquency occurring across the late teens and early twenties (see Table 2). Consistent with prior analyses utilizing these data (e.g., Wiesner et al., 2005), Elliott delinquency scores were log transformed before analysis in the following growth models (see Figure 7).

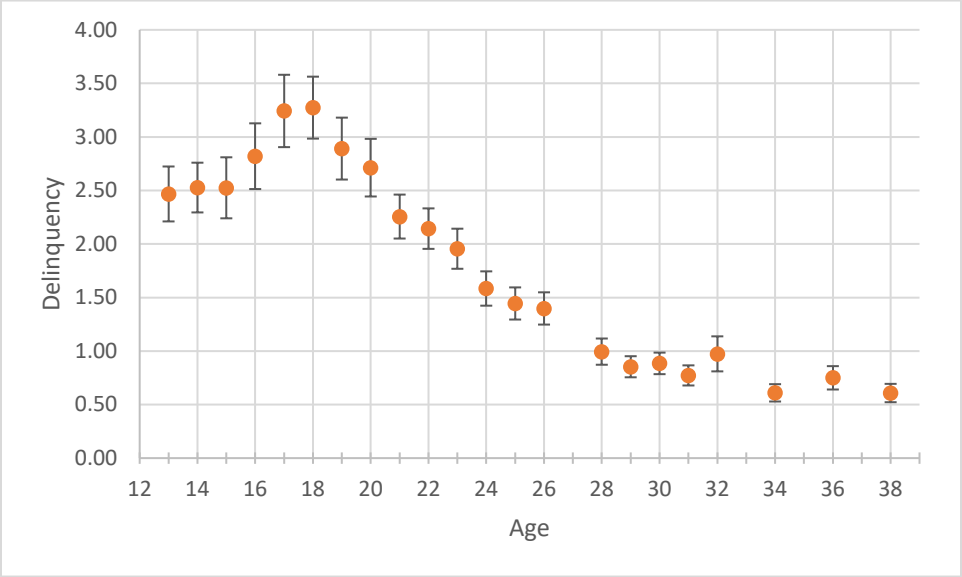


Figure 6. Mean self-reported delinquency across time. *N* ranges from 164 to 204.

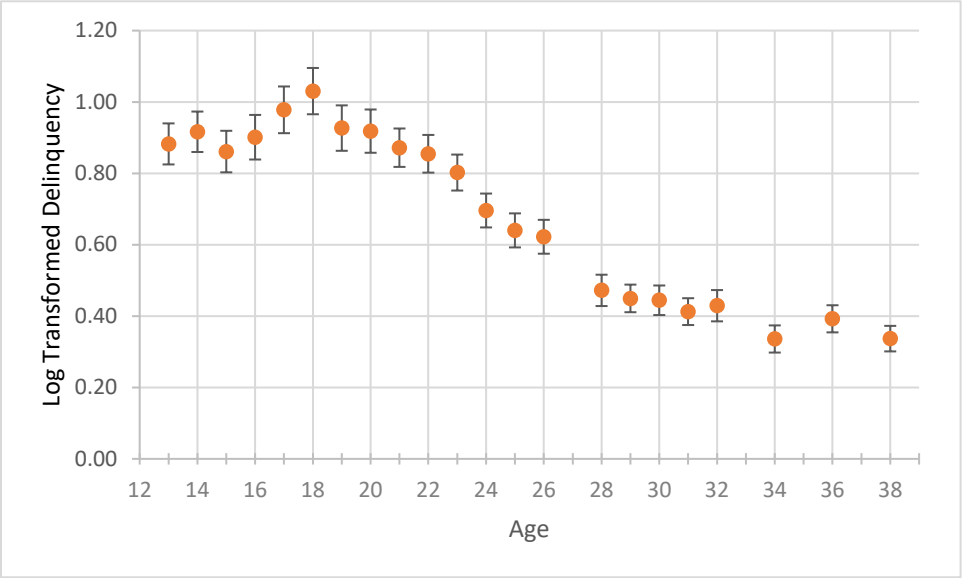


Figure 7. Mean of log transformed self-reported delinquency across time. *N* ranges from 164 to 204.

Growth Modeling

Hypothesis 1: Describing change in facial trustworthiness across ages 13 to 38.

In an exploratory hypothesis, perceived facial trustworthiness was expected to decline across ages 13 to 38, such that participants would be perceived as progressively less trustworthy as they aged (H1.1). Furthermore, such declines were predicted to temper or stabilize in adulthood, such that changes in facial trustworthiness would be less extreme or nonexistent post-adolescence (H1.2). Finally, it was predicted that there would be significant variance in intercept and slope factors (H1.3). Such individual variation in the initial levels and rate of change across time for facial trustworthiness would allow for the possibility of Dorian Gray and/or expectancy effects.

Based on visual inspection of the shape of facial trustworthiness over time (see Figure 3), a piecewise model was explored. This model separately specified growth in two developmental epochs: adolescence (ages 13-18 years) and adulthood (ages 18, 21, 36, and 38). Age 18 was chosen as the "developmental knot," or the inflection point for slope in this model, for two reasons. First, due to the rapid physiological development characteristic of adolescence, facial features may develop and change more rapidly across ages 13 to 18 than ages 18 to 38. Second, given that photographs were taken less frequently during adulthood, selecting an inflection point after age 18 would result in three or fewer time points for estimating growth in adulthood. The final three timepoints imply a positive slope, whereas when viewed in the context of complete data, no upward trend in facial trustworthiness across adulthood is implied (see Figure 3). Therefore, utilizing age 18 as the model inflection point would permit for the reliable estimation of the growth of facial trustworthiness across development.

Data from adolescence and adulthood were modeled separately to develop a model with adequate fit for each developmental epoch. Then, the adolescence and adulthood models were combined into a piecewise model spanning ages 13 to 38. The individual model development process for each developmental epoch is described below.

Adolescence Model. An intercept only model including facial trustworthiness at ages 13, 14, 15, 16, 17, and 18 years, and smiling as a time-varying covariate, demonstrated adequate fit (see Table 3; see Table 4 for model estimates).

The addition of a linear slope term resulted in good model fit (see Table 4) based on the fit indices described in the Data Analysis section. The addition of a quadratic slope term produced a nonsignificant estimate (q = .01, SE = .01, p = .071) and did not result in improvement in model fit (χ^2 (47) = 95.89, p < .001, $RMSEA$ = <.07, CFI = .92, TLI = .91; see Table 4). Thus, the linear model was retained for further analysis.

Note that the intercept value (initial level estimated at age 13) of this model corresponded to a slightly untrustworthy appearance on the 1-7 facial trustworthiness scale (3.48; see Table 4), reflecting a latent "average" facial trustworthiness score when participants were not smiling (i.e., smile rating = 1), which is somewhat lower than the raw trustworthiness rating for age 13 (M = 3.84). The effect of smiling on facial trustworthiness ratings ranged from β_{smile} = .252 to .424, all ps < .001.

Note that the negative direction of the slope indicated that on average, facial trustworthiness decreased over time (see Table 4). Variance in both the intercept (*estimate* = .11, p < .001) and slope (*estimate* = .01, p = .01) terms were statistically significant. Thus, there was evidence that initial levels of facial trustworthiness differed

across participants, and that rate of change in facial trustworthiness across adolescence also varied across participants.

Adulthood Model. Age 18 facial trustworthiness was included in both this model and the adolescence model to permit the two models to be linked at the next step. An intercept-only model including ages 18, 21, 36, and 38 demonstrated adequate model fit (see Table 3 for fit indices). Variance in the intercept term was statistically significant (see Table 4), indicating that facial trustworthiness at age 18 varied across participants after controlling for smiling. The effect of smiling on facial trustworthiness ratings was statistically significant at each wave, β_{smile} = .29 - .39, all *p*s < .001. The addition of a linear growth term to the model marginally impacted model fit as evidenced by the fit indices presented in Table 3. This model produced a nonsignificant negative slope, without statistically significant variance in the slope term (see Table 4), indicating that on average facial trustworthiness did not change across adulthood and that there was little if any variance in growth across participants. Thus, the intercept-only model best represented these data.

Piecewise Model. Linear models for adolescence (ages 13-18 years) and adulthood (ages 18-38 years)[2] defined above were combined in a piecewise model. The intercept was defined at age 13, with adolescent slope reflecting change across time from age 13 to age 18. The developmental knot (i.e., the point between two developmental epochs when slope shifts in a piecewise model) was placed at age 18, such that adulthood

[2] Although the intercept-only model was the most parsimonious fit for adulthood data, Mplus statistical software does not estimate a piecewise model in which one developmental epoch does not include a slope term. Therefore, the linear specification of the adulthood model was utilized for the piecewise model.

slope reflects change across time from age 18 to age 38. Given that the effect of smiling on facial trustworthiness was not a primary focus of this study, the effect of smiling (β_{smile} = .23 - .42, all *p*s <.001) was constrained to be equal at each wave of assessment to reduce the number of parameters estimated by the model (β_{smile} = .33, SE = .013, p < .001). This piecewise model demonstrated good model fit (see Table 3 for fit indices).

The model intercept (initial levels at age 13) was 3.46, corresponding to a slightly untrustworthy appearance on the 1-7 trustworthiness scale (see Table 4). Slope during adolescence was estimated at -.07 (p < .001), indicating that facial trustworthiness decreased across ages 13 to 18. The estimated slope parameter for adulthood was less than <.001 (p = .850), indicating that on average, there was little change in facial trustworthiness from ages 18 to 38. Note that in the final piecewise model, variance was statistically significant for all three parameters—that is, intercept, slope during adolescence, and slope during adulthood (see Table 4).

Parent Income. Parent income may be a confound variable associated with both facial trustworthiness and delinquency. To determine the need for parent income as a control variable in the following models, initial levels (at age 13) and rate of change in facial trustworthiness were regressed on parent income in the piecewise model. Parent income was not significantly associated with initial levels of facial trustworthiness ($\beta_{pincome}$ = .04, *SE* = .04, p = .242), or rate of change in facial trustworthiness during adolescence ($\beta_{pincome}$ = -0.004, *SE* = .009, p = .692) or adulthood ($\beta_{pincome}$ = -0.003, *SE* = .002, p = .212). Therefore, parent income was not retained as a control variable in the final model.

Summary of Tests of H1. Change in facial trustworthiness across time in the final piecewise model was consistent with exploratory hypotheses: that is, facial trustworthiness decreased across time (in particular, across adolescence; H1.1), those decreases stabilized during adulthood (H1.2), and variance was significant for both initial levels of facial trustworthiness at age 13 and rate of change across development, such that the intercept and slope varied across individuals (H1.3; see Table 4, see Figure 8).

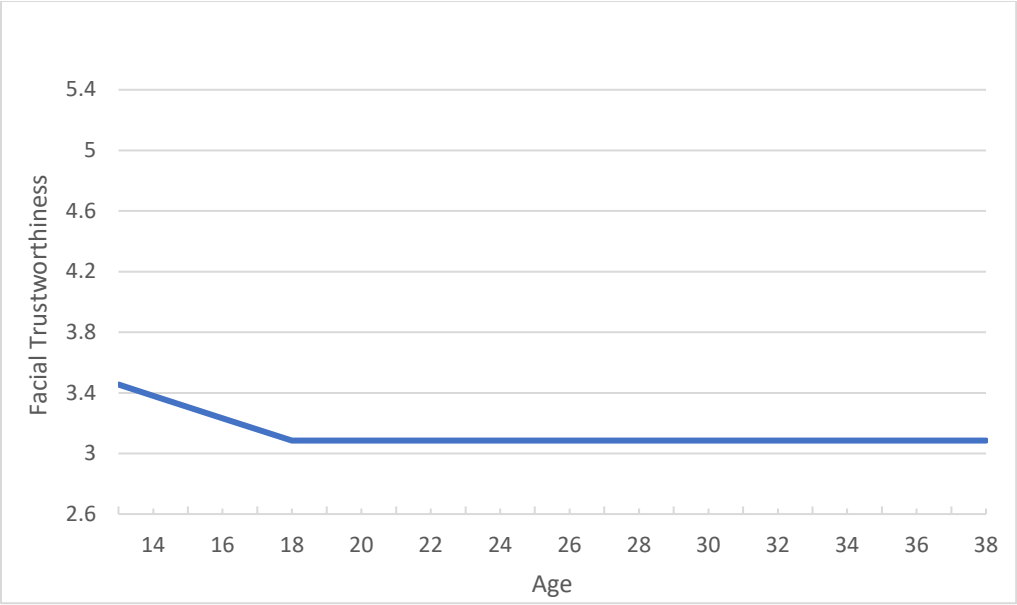


Figure 8. Final piecewise model representing growth in facial trustworthiness ages 13 to 38 years. See Table 4 for model parameters. The Y axis was set to reflect the actual range of participants' raw scores (an average of *k* raters' ratings; see Table 1), rather than the full range used by individual raters (1 to 7).

Table 3. *Fit indices for facial trustworthiness growth models.*

Model	χ^2	DF	*p*	RMSEA	CFI	TLI
Adolescence-only model						
Intercept only	111.52	49	<.001	0.079	0.90	0.89
Linear	60.82	46	.070	0.040	0.98	0.97
Adulthood-only model						
Intercept	63.48	20	<.001	0.10	0.87	0.86
Linear	57.89	17	<.001	0.11	0.88	0.84
Piecewise (adolescence and adulthood) model						
Unconstrained smile	167.44	108	<.001	0.05	0.93	0.93
Constrained smiling	232.94	116	<.001	0.07	0.87	0.87

Note: Adequate model fit is indicated where *p* of χ^2 (df) ≥ .05, *RMSEA* = <.08, *CFI* ≥ .90, *TLI* ≥ .90. Good model fit is indicated where χ^2 (df) is nonsignificant, *RMSEA* < .05, *CFI* ≥ .95, *TLI* ≥ .95.

Table 4. *Unstandardized estimates for facial trustworthiness growth models*.

	Intercept					Slope				
Model	Est.	SE	*p*	Variance	*p*	Est.	SE	*p*	Variance	*p*
Adolescence-only model										
Intercept	3.26	.03	<.001	.091	<.001	-	-	-	-	-
Linear	3.48	.04	<.001	.114	<.001	-0.08	.01	<.001	.005	.012
Adulthood-only model										
Intercept	3.06	.03	<.001	.057	<.001	-	-	-	-	-
Linear	3.02	.05	<.001	.091	.003	.003	<.01	.263	<.001	.054
Piecewise (adolescence and adulthood) model										
Adolescent	3.46	.04	<.001	.109	<.001	-0.074	.01	<.001	.003	.043
Adult	-	-	-	-	-	<.001	<.01	.850	<.001	.005

Hypothesis 2: Interrelating facial trustworthiness and delinquency across time.

Delinquency Model. As the shape of delinquency across time is not the primary theoretical focus of the current project, the definition of the delinquency model is described in less detail than that of the facial trustworthiness growth model.

Following the same methodology described under analyses for Hypothesis 1, an intercept only model for delinquency during adolescence (ages 13-18) was constructed. This model demonstrated adequate fit; however, the introduction of a linear term

improved model fit as evidenced by fit indices (see Table 5). The linear model produced a positive slope (see Table 6), indicating that on average, delinquency increased over adolescence. A linear model during adulthood (ages 18-38) indicated a negative linear slope, such that delinquency decreased across time (see Table 6); however, this model demonstrated poor fit (see Table 5). In a quadratic model, a small, positive quadratic term emerged as statistically significant (q_{del} = .002, *SE* <.001, $p < .001$), indicating that over adulthood, delinquency decreased less rapidly as participants aged. The quadratic model demonstrated good fit, compared to the fit indices of the previous adulthood delinquency models (see Table 5). Because the scaling of delinquency resulted in a variance estimate prohibitively small for Mplus to estimate, variance of the quadratic term was constrained to zero.

Combining the linear adolescence model and the quadratic delinquency model resulted in a final piecewise model that demonstrated adequate fit (see Table 5; Figure 9). A positive, linear slope (β = .036) during adolescence indicated that delinquency increased from ages 13 to 18. The negative linear (β = -.071) and positive quadratic (q_{del} = .002, p < .001) terms during adulthood indicated that delinquency decreased across ages 18 to 38, and that those decreases became less rapid and pronounced over time (see Table 6). The differential growth trajectories across developmental epochs were important to consider when interpreting the hypothesis tests described below.

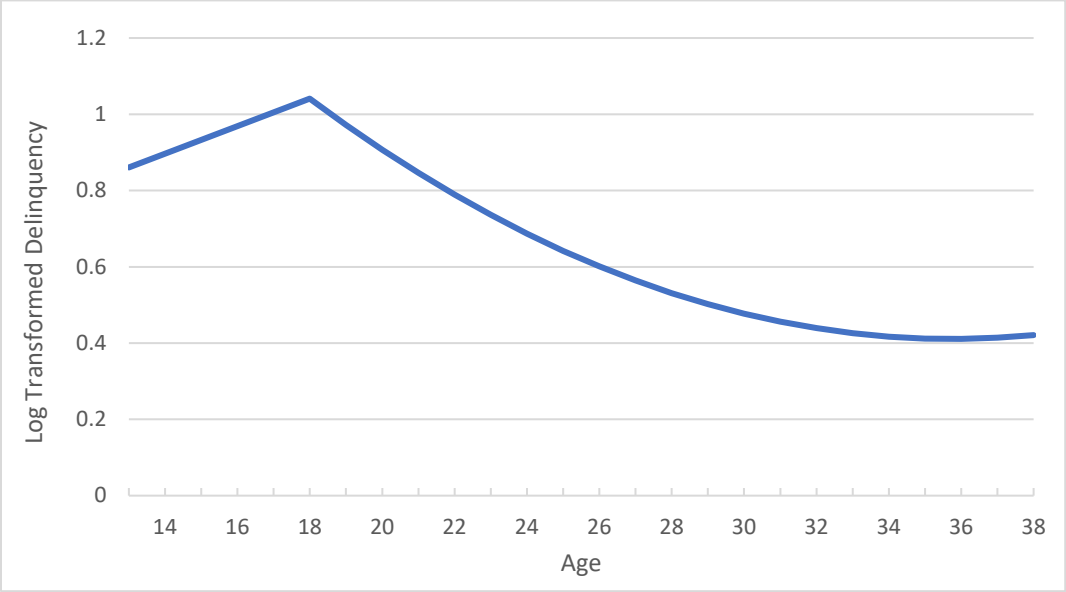


Figure 9. Final piecewise model representing growth in log transformed delinquency ages 13 to 38 years. See Table 6 for model parameters.

Table 5. *Fit indices for delinquency growth models.*

Model	χ^2	DF	*p*	RMSEA	CFI	TLI
Adolescence-only model						
Intercept	45.63	13	<.001	0.11	0.93	0.95
Linear	31.50	16	.012	.07	0.98	0.98
Adulthood-only model						
Intercept	1057.98	151	<.001	0.17	0.50	0.55
Linear	382.10	148	<.001	0.09	0.87	0.88
Quadratic	315.39	147	<.001	<.001	0.91	0.91
Piecewise (adolescence and adulthood) model						
Full model	603.11	243	<.001	.085	.85	.86

Note: Adequate model fit is indicated where *p* of χ^2 (df) ≥.05, *RMSEA* = <.08, *CFI* ≥ .90, *TLI* ≥ .90. Good model fit is indicated where χ^2 (df) is nonsignificant, *RMSEA* < .05, *CFI* ≥ .95, *TLI* ≥ .95.

Table 6. *Unstandardized estimates for delinquency growth models.*

	Intercept					Slope				
Model	Est.	*SE*	*p*	Variance	*p*	Est.	*SE*	*p*	Variance	*p*
Adolescence-only model										
Intercept	.92	.05	<.001	.420	.047	-	-	-	-	-
Linear	0.87	.05	<.001	.427	<.001	0.03	.01	.046	.018	<.001
Adulthood-only model										
Intercept	0.54	.03	<.001	.175	.019	-	-	-	-	-
Linear	0.92	.05	<.001	.410	.046	-0.040	<.01	<.001	.001	<.001
Quadratic	1.07	.05	<.001	.412	<.001	-0.077	.01	<.001	.001	<.001
Piecewise (adolescence and adulthood) model										
Adolescent	0.86	.06	<.001	.537	<.001	0.036	.013	.006	.024	<.001
Adult	-	-	-	-	-	-0.071	.005	<.001	.001	<.001

Parallel Process Model. The piecewise growth models for facial trustworthiness and delinquency were combined into a single parallel process model ($\chi^2(746) = 1271.19$, $p < .001$). Intercepts and time-concurrent slopes were allowed to covary across processes (i.e., adolescent slopes for delinquency and facial trustworthiness were allowed to covary with one another, and adulthood slopes were allowed to covary with one another; see Figure 10).

Parameter estimates were consistent with the prior models run separately for facial trustworthiness and delinquency (see Table 7): The intercept for facial trustworthiness at age 13 reflected a score slightly below the center of the 1-7 facial trustworthiness scale, the negative adolescence slope indicated that on average, facial trustworthiness decreased across ages 13 to 18 (controlling for smiling); the adulthood slope failed to detect significant change in facial trustworthiness across ages 18 to 38. There was statistically significant variance in the intercept and slope terms for the facial trustworthiness model, indicating that participants started at different levels of facial trustworthiness, and that growth across time varied between participants. The positive

slope term for delinquency during adolescence and negative slope term during adulthood indicated that on average, delinquency increased across ages 13 to 18 and decreased across ages 18 to 38. Given that these data were log transformed, the interpretation of the intercept term for delinquency was not meaningful. Back-transforming the estimated intercept (exponentiating .86) yielded a score of 2.36, indicating that on average, participants reported engaging in roughly two delinquent acts at age 13.

Table 7. *Unstandardized estimates for full piecewise parallel process model for delinquency and facial trustworthiness.*

	Intercept					Slope				
Full Model	Est.	*SE*	*p*	Variance	*p*	Est.	*SE*	*p*	Variance	p
Delinquency										
Adolescence	0.86	0.06	<.001	.538	<.001	0.04	0.01	.006	.024	<.001
Adulthood	-	-	-	-	-	-0.07	0.01	<.001	.001	<.001
Facial Trustworthiness										
Adolescence	3.45	0.04	<.001	.108	<.001	-0.07	0.01	<.001	.003	.045
Adulthood	-	-	-	-	-	<0.01	<0.01	.863	<.001	.005

Note: Smiling is utilized as a time-varying control variable for each observation of facial trustworthiness.

After fitting the model, four regression paths were introduced to test primary hypotheses. That is, the intercept of delinquency was set as a predictor for the slopes of facial trustworthiness during adolescence and adulthood (H2.1, Dorian Gray effects), and the intercept of facial trustworthiness was set as a predictor for the slopes of delinquency during adolescence and adulthood (H2.2, expectancy effects; see Figure 10).

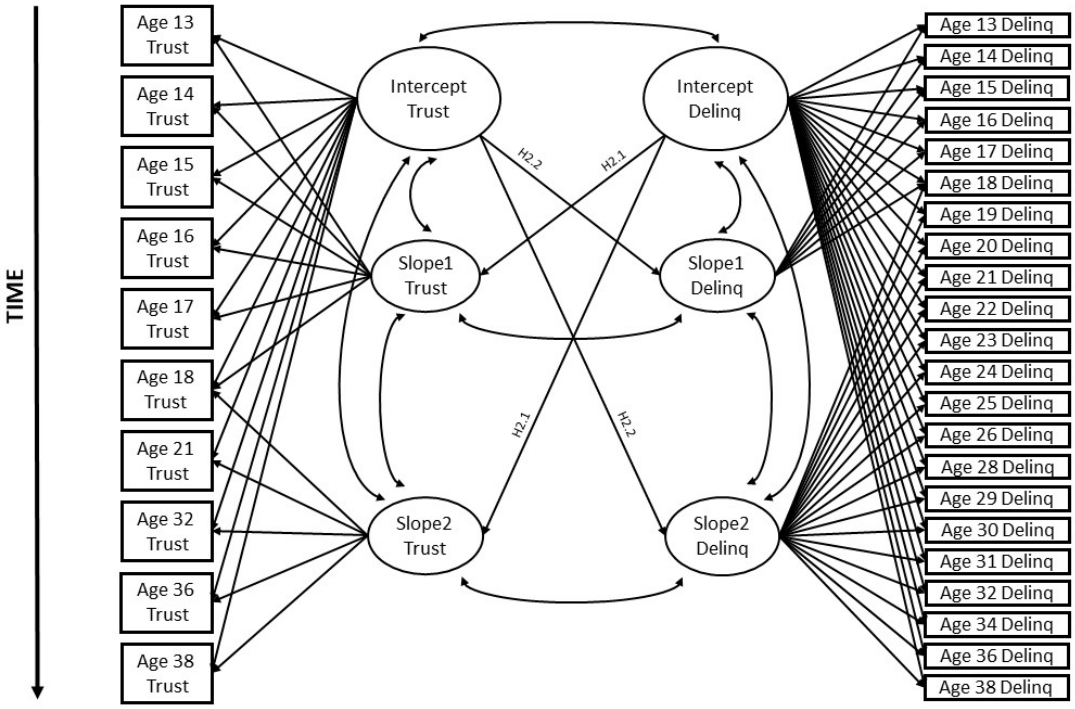


Figure 10. A graphical representation of the piecewise parallel process model. Smile control variables and the quadratic term for delinquency not depicted.

H2.1: Dorian Gray Effects. As reviewed above, engagement in delinquency may lead to the development of an untrustworthy appearance congruent with those behaviors—a Dorian Gray effect. It was hypothesized that initial levels of delinquency in this sample may be associated with decreases in facial trustworthiness across time, consistent with the Dorian Gray effect. To test this hypothesis, initial levels of delinquency at age 13 (intercept) were tested as a predictor of slopes of facial trustworthiness during adolescence and adulthood.

Adolescence Epoch in Piecewise Model. In the context of decreases in facial trustworthiness across adolescence, adolescents who had higher initial levels of delinquency experienced greater decreases in facial trustworthiness across time than

those with lower initial levels of delinquency. That is, consistent with study hypotheses, delinquency and facial trustworthiness were negatively associated across adolescence, such that facial trustworthiness decreased more among those who engaged in more delinquent behaviors initially (β = -.033, *SE* = .016, *p* = .036).

Adulthood Epoch in Piecewise Model. A nonsignificant, negative slope was estimated for rate of change in facial trustworthiness across time during adulthood; that is, on average, there was little or no change detected in participants' facial trustworthiness across ages 18 to 38. However, there was significant variation across individuals, such that rate of change varied between participants.

Accounting for the relationship between intercept of delinquency and growth in facial trustworthiness during adolescence, initial levels of delinquency were not associated with changes in facial trustworthiness during adulthood (β = -.003, *SE* = .003, *p* = .389). Thus, contrary to study hypotheses, there was no evidence that delinquency at age 13 was associated with decrements in facial trustworthiness during adulthood.

H2.2: Expectancy Effects. An untrustworthy appearance may elicit suspicion and negativity from social partners, leading an untrustworthy appearing person to develop untrustworthy behaviors over time—i.e., expectancy effects. It was hypothesized that initial levels of facial trustworthiness in this sample may be associated with rate of change in delinquency across time in this sample, consistent with expectancy effects. To test this hypothesis, initial levels of facial trustworthiness at age 13 (intercept) were set as a predictor of rate of change in delinquency during adolescence and adulthood (slope).

Adolescence Epoch in Piecewise Model. In the context of increasing delinquency across adolescence, participants who had higher initial levels of facial trustworthiness

demonstrated reduced escalation in delinquency, compared to their less trustworthy appearing peers. That is, consistent with study hypotheses, initial levels of facial trustworthiness were negatively associated with delinquency across adolescence, such that delinquency increased less among those who looked more trustworthy at age 13 (β = -.893, *SE* = .353, *p* = .011).

Adulthood Epoch in Piecewise Model. In the context of declining delinquency across adulthood, and accounting for the association between intercept of facial trustworthiness and slope of delinquency during adolescence previously described, participants who had higher levels of facial trustworthiness at age 13 decreased in delinquency more rapidly than their peers who looked less trustworthy across adulthood. That is, consistent with study hypotheses, initial levels of facial trustworthiness were negatively associated with changes in delinquency from 13 to 18 years, such that delinquency decreased more among those who looked more trustworthy at age 13 (β = -.120, *SE* = .051, *p* = .019).

Discussion

A person's appearance can dramatically influence the nature and consequences of social interactions (Todorov et al., 2015). Facial trustworthiness, read by perceivers as a signal to a target's approachability, is a particularly influential dimension of appearance for interpersonal evaluations and consequent social interactions (Oosterhof & Todorov, 2008). The present study investigated the development of facial trustworthiness across time. Because delinquent behavior is powerfully influenced by social factors, especially during adolescence (Dishion, 2000; Jessor, 1991; Van Ryzin & Dishion, 2014), the present study also investigated how facial trustworthiness might relate to engagement in

delinquent behaviors across the lifespan. In particular, the extent to which facial appearance predicted the development of delinquency across time (expectancy effects), and the extent to which engagement in delinquent behaviors predicted later changes in facial trustworthiness (Dorian Gray effects), were assessed. To address these questions, prospective self-reports of delinquency from an at-risk sample of 206 boys were related to thin-slice ratings of facial trustworthiness across their development.

Main study findings will be summarized here, and then unpacked in the subsequent sections of the discussion. In terms of developmental trends, facial trustworthiness was found to decrease over time. Since a child generally has less capacity to do harm than an adult, this is consistent with prior research highlighting that facial trustworthiness is associated with perceptions of approachability versus threat (Oosterhof & Todorov, 2008; Todorov et al., 2008). No evidence of changes in facial trustworthiness was found after adolescence and into adulthood, perhaps because adulthood is not characterized by the same rapid physiological facial development experienced during adolescence (Marečková et al., 2011).

Regarding the interrelationships between facial trustworthiness and delinquency, two intriguing effects were observed. Consistent with expectancy effects, levels of facial trustworthiness at the study outset were associated with later engagement in delinquency. Consistent with Dorian Gray effects, initial levels of delinquency were associated with later decrements in facial trustworthiness. Each of these effects, their interpretations, and possible confounds are discussed below.

Development of Facial Trustworthiness

The present study was the first to investigate the development of facial trustworthiness across adolescence and adulthood. Initial levels, growth across time, and variance between participants in facial trustworthiness were examined among this sample of 206 at-risk men followed from ages 13 to 38. As hypothesized, facial trustworthiness decreased over adolescence, such that on average, participants appeared more trustworthy at age 13 than they did at age 18. There were significant individual differences in initial levels of facial trustworthiness at age 13, and furthermore, there was significant variance in growth of facial trustworthiness across time between participants. However, during adulthood, there was no evidence of further developments in facial trustworthiness.

These results indicate that developmental trends in facial trustworthiness vary across persons and on average, facial trustworthiness decreases over time. In a prior study, babyfacedness, a facial feature associated with perceptions of trustworthiness, was also shown to decrease among men over the lifespan (Zebrowitz et al., 1993); yet, there was no evidence of the stabilization across ages 18 to 38 apparent in the present study for facial trustworthiness. Adolescence may be a critical period for the development of a trustworthy or untrustworthy appearance. If behavior can influence facial trustworthiness (e.g., via tobacco use; Alley et al., 2019), those influences may be most consequential during adolescence and readily apparent by early adulthood, during which the consequences of early decrements in facial trustworthiness may interfere with the declines in delinquency characteristic of this developmental period (Sampson & Laub, 2003; Wiesner et al., 2005).

Dorian Gray Effects

In the context of a general trend of decline from ages 13 to 18, facial trustworthiness decreased more among those with higher initial levels of delinquency. That is, adolescents who engaged in more delinquency at the study outset experienced the greatest declines in facial trustworthiness up to age 18. These results are consistent with the possibility that behavior might influence facial appearance in a manner consistent with Dorian Gray effects.

Although initial levels of delinquency were associated with rate of change in facial trustworthiness during adolescence, this effect was not observed regarding development of facial trustworthiness during adulthood. This may represent the temporal disconnect between when facial trustworthiness was measured (age 13) and the developmental window assessed (ages 18-38), such that engagement in delinquency during adulthood may be more important for development of facial trustworthiness across adulthood than engagement in delinquency during adolescence. It may also reflect the fact that there was little growth across time in facial trustworthiness after age 18; i.e., features of appearance that influence this perception may be less malleable after adolescence.

Thus, adolescents who engage in more delinquent behaviors early in life may develop a more untrustworthy appearance by early adulthood. The harsher social environment associated with low facial trustworthiness (e.g., Q. Li et al., 2017; Wilson & Rule, 2015, 2016; Wu et al., 2018) could interfere with social relationships, educational opportunities, or employment prospects, perhaps interfering with the desistence processes characteristic of early adulthood (Giordano et al., 2003; Runell, 2017; Sampson & Laub,

2003; Skardhamar & Savolainen, 2014; Wiesner et al., 2005). Furthermore, given the limited variance in facial trustworthiness growth observed during adulthood, there is no evidence that facial trustworthiness might improve later, even if delinquency desisted.

Although parent income, a potential confounding variable, was not associated with study outcomes, it still is possible that other unmeasured factors might have led to both engagement in delinquency and relatively greater degradations in facial trustworthiness across adolescence. Still, these findings are consistent with the Dorian Gray hypothesis, or the idea that delinquent behavior may lead to a less trustworthy appearance.

Expectancy Effects

In the context of escalating engagement in delinquency during adolescence, adolescents who looked less trustworthy at age 13 experienced more rapid escalation in delinquency than their peers who looked more trustworthy. That is, those who looked untrustworthy became involved in delinquency more rapidly and desisted in delinquency more slowly across adolescence and adulthood. This is consistent with expectancy effects; i.e., those who appear untrustworthy may elicit a social environment that increases the likelihood that they behave in untrustworthy ways. Important social partners (e.g., peers, teachers, employers, etc.) may interpret the behavior of an untrustworthy appearing target negatively and respond to them with suspicion, much as teachers in Rosenthal's Pygmalion study (Rosenthal & Jacobson, 1968) responded differently to students whom they were erroneously led to believe had unique potential to excel.

Indeed, prior work has found that children with higher levels of facial trustworthiness were more popular among their peers at school than those who were not. Furthermore, those children were perceived by their peers to behave in more trustworthy ways over time, which was mediated by the increases in popularity associated with a trustworthy face (Q. Li et al., 2017). Although Li and colleagues measured trustworthy behavior differently than the present study (peers' report of the target's ability to keep a promise or a secret, versus self-report of delinquency) and over a shorter developmental window, both studies are similar in that they demonstrate untrustworthy-type behavior following an untrustworthy facial appearance in a developmental context.

In the context of deescalating delinquency during adulthood, levels of facial trustworthiness during early adolescence were associated with greater de-escalation, such that those who appeared more trustworthy during early adolescence decreased more in delinquency than those with less trustworthy appearances. It is remarkable that strangers' impressions of participants' appearances at age 13 predicted rate of change in delinquency between ages 18 to 38, two decades later. If, as hypothesized, expectancy effects do underlie the relationship between facial trustworthiness and patterns of delinquency over adolescence, these results are evidence that such a relationship may persevere well into adulthood.

Effect of Smiling

Perceivers' perceptions of facial trustworthiness are higher when the target is smiling (Krumhuber et al., 2007; Ozono et al., 2010). It was therefore necessary to control for the effect of smiling on perceptions of facial trustworthiness at each wave of assessment, as some participants smiled during some photograph and not others.

Although the homogeneity of the current sample in terms of race and socioeconomic status limits generalizability, given the numerous repeated measures across a wide developmental window, the present study may have been the most powerful estimate of the effect of smiling on perceptions of boys' and men's facial trustworthiness yet conducted.

Consistent with prior literature, smiling had a strong, positive relationship with perceptions of trustworthiness. This was true for the men in this sample at all ten ages for which photographs were available, ranging from early adolescence to middle adulthood. These findings are consistent with prior research (Krumhuber et al., 2007; Ozono et al., 2010) and the theoretical perspective that perceptions of trustworthiness drawn from the face are an overgeneralization of emotional cues to approachability (e.g., smiling versus scowling; Oosterhof & Todorov, 2008); i.e., the emotion-overgeneralization hypothesis (Zebrowitz & Montepare, 2006). Thus, just as a trustworthy appearing target may have a face reminiscent of a smile, a perceivers' assessment of a target's trustworthiness will naturally be higher when the target actually is smiling.

Strengths, Limitations, and Future Directions

Study Strengths

The present study had several key strengths. First, the longitudinal design allowed for prospective assessment of participants across two decades, minimizing biases due to retrospective reporting. Photographs were taken annually during adolescence, a period characterized by pronounced physiological development, and then assessment continued less frequently into adulthood. Appearance and delinquency were assessed

prospectively across time, allowing for the estimation of interrelations of delinquency and facial trustworthiness across two developmental epochs.

A second study strength was that raters naïve to the study purpose and the method of participants' recruitment rated the photographs for trustworthiness. That is, raters did not know that the participants were from an at-risk sample, or that participants' appearances would be examined in relation to their history of delinquency. The fact that impressions of the participants' trustworthiness made by strangers three decades after the earliest photographs were taken were associated with participants' trends in delinquency speaks to both the stability across time and cross-cutting influence of impressions of trustworthiness.

Another study strength was the nature of the sample. Given that participants were initially recruited from neighborhoods with higher than average police-reported juvenile delinquency, delinquency was relatively common in this sample. Theoretically, Dorian Gray and expectancy effects may be most pronounced and influential among at-risk groups, for whom there would be more opportunities to engage in delinquency. Thus, the present sample allowed for the investigation of these issues among a group who may be disproportionately affected by them. Practically, the frequency with which delinquent acts were reported allowed for the analysis of growth across time in delinquency in relation to facial trustworthiness, and growth in facial trustworthiness in relation to delinquency, which may have been more difficult or impossible in a community sample.

Limitations and Future Directions

The present study had several key limitations. The composition of the sample was largely white (90%), so it is unclear to what extent these effects generalize to racially

and ethnically representative samples. Given the low frequency of nonwhite participants, it was impossible to examine whether ethnicity moderated the effects observed, yet it is known that a person's race has a powerful influence on interpersonal judgments about them (e.g., King & Light, 2019; Stanley et al., 2011). Furthermore, all participants were male, so it is unknown whether these patterns would generalize to girls and women. Indeed, there is evidence that facial structure may be more associated with men's behavior than women's behavior (Foo et al., 2019), perhaps indicating that expectancy effects related to facial trustworthiness may be different for women and men. Therefore, future research should replicate these analyses with more diverse samples, to 1) better represent women and other ethnic groups and 2) allow for the investigation of interactions between ethnicity, gender, and facial trustworthiness.

Although a primary interest in this study was to determine growth in facial trustworthiness across time, it should be noted that faces were rated within wave. That is, raters viewed all faces in a given wave (e.g., age 13) and rated them for trustworthiness, and any individual rater did not see faces from other age groups. This may result in anchoring for raters, such that faces may have been rated as trustworthy *in relation to the other faces seen*, rather than trustworthy in relation to all possible faces of all possible ages (Todorov et al., 2015). If raters viewed faces of all ages when rating trustworthiness (as was done by Zebrowitz et al., 1993), the discrepancy between older and younger faces' trustworthiness that emerged in the present analysis may have been yet more dramatic and better represented the development of facial trustworthiness across time.

Another set of study limitations regard factors that went unmeasured. For example, it remains unclear by what mechanisms delinquency could lead to a less

trustworthy appearance. It is possible that engagement in violent or deviant groups and activities elicits more negative facial expressions, leading to changes in wrinkling or musculature of the face that resemble a scowl at rest. It is also possible that such affiliation might foster grooming behaviors that lead to a less trustworthy appearance. Finally, especially during adolescence, delinquency and substance use are closely linked (Jessor, 1991; Van Ryzin & Dishion, 2014), and there is some evidence that tobacco use (which covaried with delinquency in this sample) might itself lead to decrements in facial trustworthiness (see Alley et al., 2019). Thus, substance use may have mediated the observed Dorian Gray effects, such that adolescents who engaged in more delinquency used more substances. Alternatively, the observed relationship between facial trustworthiness and delinquency may have been spurious and entirely driven by substance use. The complexity of the piecewise parallel process model precluded the inclusion of tobacco or other substance use in the present study. Future research should investigate the role that tobacco use, facial expression, and deviant peer affiliation may play in contributing to or mediating the effects observed in the present study.

Second, the present study did not measure the intermediary processes involved in expectancy effects. Expectancy effects in this context were theorized to result when a target who appears untrustworthy elicits a negative response from perceivers that constrains the target's behavioral opportunities, such that the target's behaviors match the perceivers' expectations. Although data regarding participants' appearance and participants' behavior were collected, perceivers' attributions and interpretations of the target's behavior were theorized but not measured in the present study. The two-decade developmental window and the observational nature of this study were not best suited to

collecting data on the microsocial interactions that underlie expectancy effects. However, laboratory research should investigate how an untrustworthy appearance may affect perceivers' attributions of a target's behavior in ways that may have cumulative impacts on the target across time.

Conclusion

In a sample of 206 at-risk men followed from ages 13 to 38, facial trustworthiness declined from early adolescence to age 18 and then stabilized from ages 18 to 38. Consistent with expectancy effects, individual differences in initial levels of facial trustworthiness at age 13 were associated with increased engagement in delinquency during adolescence and adulthood in a parallel process model. Consistent with Dorian Gray effects, initial levels of delinquency predicted development of facial trustworthiness across adolescence, such that those who engaged in more delinquency decreased more in facial trustworthiness than their peers who had engaged in less. The present study was the first analysis of the development of facial trustworthiness across adolescence and adulthood. These results provide strong evidence that facial trustworthiness and delinquency are interrelated across the development of boys and men.

Study 2

Ambiguous Behavioral Information and Implicit Biases Related to Facial Trustworthiness

Despite the common adage "don't judge a book by its cover," people make rapid and implicit evaluations of other's personalities based only on their appearances (Todorov et al., 2015). For example, people with rounder faces, high curved brows, and upturned lips are perceived as more trustworthy than those with square faces, low slanted brows, and downturned lips (see Figure 11; Todorov et al., 2008). These evaluations are consistent among perceivers and targets of various ages and ethnicities (Birkás et al., 2014; Charlesworth et al., 2019; Q. Li et al., 2017) and reliable across perceivers within milliseconds of exposure to a face, such that increasing exposure time does not significantly alter those evaluations (Todorov et al., 2009). Subliminal exposure to a trustworthy face results in more positive attitudes toward a formerly neutral stimulus (Shen et al., 2020), and exposure to an untrustworthy face is associated with activation of the amygdala (Todorov et al., 2008). As noted by Todorov and colleagues (2015), evaluations based on facial characteristics are inevitable, because they are less to do with conscious thought than they are with the process of perception. Such implicit appearance-based judgments have the power to shape social interactions, with the potential for quite negative consequences for those who appear untrustworthy.

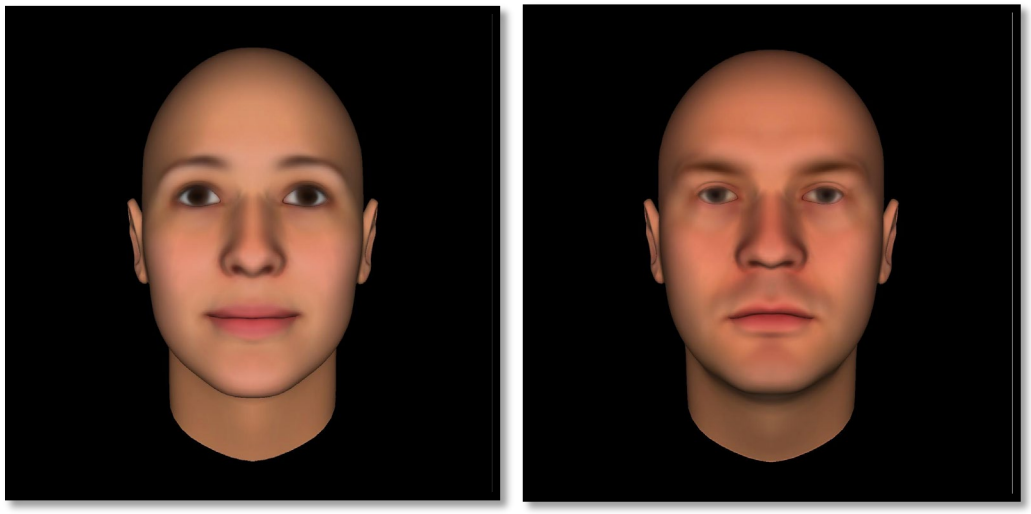


Figure 11. A computer-generated prototypically trustworthy (left) and untrustworthy (right) face, based on models by Oosterhof and Todorov (2008).

Perceivers' Implicit Trustworthiness Evaluations Affect their Behavior

Implicit appearance-based judgments may influence the social environment of the person being perceived (i.e., the target) by influencing the perceiver's behavior. For example, participants' judgments of another person's likeability based on a facial photograph predicted how warmly they approached the person one month later, and furthermore, how much they rated actually liking that person after meeting (Gunaydin et al., 2017). Likewise, there is evidence that targets' facial trustworthiness may affect a perceiver's behavior toward the target.

For example, during a lab-based trust game in which participants had the opportunity to lend money to a partner who could cooperate or cheat (this 'partner' was actually an algorithm paired with a picture of a face), participants invested more money with partners who looked trustworthy than those who did not (van 't Wout & Sanfey, 2008). In a similar paradigm, economic incentives for a partner to cheat were weighed less heavily in participants' investment decisions than facial trustworthiness, even though

participants reported believing that those incentives were more valid predictors of behavior than appearance (Jaeger et al., 2019). In another lab task, participants required less evidence before they arrived at a guilty verdict and were more confident in the guilty verdict when the defendant's face was untrustworthy (Porter et al., 2010). Furthermore, when allowed to deny themselves a reward in order to punish a partner who treated them unfairly, participants were more likely to punish targets who looked less trustworthy (Wu et al., 2018).

These effects observed in the laboratory translate to real-world court case decisions, where men whose faces were rated as untrustworthy by naïve raters were more likely to have received the death sentence—even among those who were later exonerated (Wilson & Rule, 2015, 2016). Facial trustworthiness even affects children's behavior; by five years old, children not only judge trustworthy-looking people more favorably than those who look less trustworthy, but they are also more likely to give them gifts (Charlesworth et al., 2019). Thus, facial trustworthiness appears to shape the way that perceivers initially respond to a target and to exacerbate perceivers' negative reactions to evidence that the target engaged in antisocial behavior. It is therefore imperative to determine the extent to which these judgments are modifiable in the face of counter-evidence; that is, whether it is possible to attenuate a potentially erroneous negative evaluation based on an untrustworthy face.

Modifying Implicit Evaluations Based on Facial Trustworthiness

General theories of person perception indicate that perceivers consider "bad" behavior to be more diagnostic of a target's true character than "good" behavior (Schaller, 2008). For example, a person who volunteers in a soup kitchen every Sunday

but once tortured an animal for pleasure would be considered "bad", even though the "bad" behavior occurred less frequently than the "good" behavior. Indeed, this diagnostic asymmetry has been replicated in a laboratory environment, in which perceivers were more likely to reappraise their evaluation of a target based on new behavioral information when the perceiver's initial impression was positive, rather than negative (Cone & Ferguson, 2015).

There is strong evidence that facial trustworthiness affects the manner in which perceivers approach and initially evaluate a target. Consistent with the theory of diagnostic asymmetry, it seems likely that perceivers' positive evaluations of a trustworthy-appearing target may be easier to manipulate than their negative evaluations of an untrustworthy appearing target. Indeed, during a trust-based economic task, participants did not advantage partners with trustworthy faces over those with untrustworthy faces when given negative behavioral information about the target (T. Li et al., 2017). Furthermore, a single piece of highly diagnostic information was capable of reversing an implicit evaluation based on facial trustworthiness (e.g., *this person tortured a defenseless animal*; Shen et al., 2020). Thus, a perceiver may reevaluate a positive evaluation of a target when told that person tortured a defenseless animal; however, that perceiver will likely still be suspicious of someone who looks untrustworthy even after hearing that that person regularly volunteers in a soup kitchen.

Although facial characteristics are poor indicators of future behavior (Todorov et al., 2015), perceivers seem to equate an untrustworthy appearance with information about that person's innate characteristics. Those evaluations are modifiable in a manner consistent with judgments based on behavioral information about the target (e.g., Cone &

Ferguson, 2015); i.e., it is easier to change a positive evaluation based on facial trustworthiness than a negative one. Although it is clear that extremely diagnostic information can influence trustworthiness judgments based on the target's facial features (Shen et al., 2020), it is less clear from the literature how ambiguous behavioral information may be interpreted in the context of a trustworthy or untrustworthy face, or how the repeated social disadvantages of an untrustworthy appearance may affect the target in the long-term.

Expectancy Effects: Long Term Consequences of an Untrustworthy Appearance

Given the social consequences of an untrustworthy appearance, it is worth asking how perceivers' negative reactions to an untrustworthy face might shape an untrustworthy-appearing person across time. Zebrowitz (1997) posits that one possibility is a self-fulling prophecy, or *expectancy effect*, whereby individuals begin to behave in a manner consistent with their facial appearance. Although trustworthiness judgments based on appearance are generally erroneous (Todorov et al., 2015), there are several examples of an accurate relationship between appearance and uncooperative or "untrustworthy" behavior. As one example, participants were able to accurately identify faces of men who reported engaging in romantic infidelity with above-chance accuracy (Foo et al., 2019). Similarly, a facial composite of participants who reported they would be likely to defect in the prisoner's dilemma was rated by perceivers as less cooperative-looking than a composite of those who reported they would cooperate (note: photographs were taken before participants made this report; Little et al., 2013). Furthermore, among children ages 8-12 years, facial trustworthiness predicted peers' assessments of the

students' ability to keep a secret or a promise one year later (Q. Li et al., 2017), which may have resulted from expectancy effects.

Taken together, there is ample evidence that 1) people make implicit evaluations about others' behavioral tendencies based on their facial trustworthiness (Shen et al., 2020; Todorov et al., 2015), 2) people interpret positive and negative behavioral information about a target differently based on how trustworthy the target appears (i.e., "good" behavioral information may be discounted if the person appears trustworthy; Shen et al., 2020; Wu et al., 2018), and 3) when two targets have engaged in the same behavior, perceivers behave differently toward someone who appears untrustworthy than they would someone who appears trustworthy (e.g., by assigning harsher punishments; Jaeger et al., 2019; Wilson & Rule, 2015; Wu et al., 2018). Therefore, people may approach an untrustworthy-appearing person with suspicion (thus eliciting negative reactions from the target), punish them harshly for minor transgressions, and ignore their positive behaviors such that the target receives less reinforcement for prosocial behavior than would a trustworthy appearing person. The cumulative effect over time would be an individual who is more likely to behave in untrustworthy ways—i.e., an expectancy effect.

Facial Trustworthiness and Ambiguous Behavioral Information

Prior work has investigated the potential for positive and negative behavioral information to modify perceivers' implicit evaluations of people based on their facial trustworthiness (e.g., Shen et al., 2020). However, there has been little to no investigation of how the target's facial trustworthiness might influence perceivers' interpretation of ambiguous behavioral information about the target (but see Porter et al.,

2010). In an ambiguous situation, perceived facial trustworthiness may cause a perceiver to give the target the benefit of the doubt or to make a negative attribution about their intentions—and these decisions may be a powerful mechanism for the development of expectancy effects.

Perceivers' evaluations of targets based on ambiguous behavioral information may be particularly susceptible to the influence of facial trustworthiness. Ambiguous behavioral information may even intensify the influence of facial trustworthiness on interpersonal evaluations. Consider the classic study by Darley and Gross (1983), in which participants were asked to estimate the academic abilities of a grade school girl. In this study, the introduction of ambiguous behavioral information dramatically influenced the effect of socioeconomic stereotypes on perceivers' evaluations of the girl's performance. When only given information about the girl's socioeconomic status (SES), participants who were led to believe she came from a low SES and those who were led to believe she came from a high SES both rated her academic abilities as roughly at her grade level. A subset of participants were shown a video clip of the child performing a verbal test. Key to this study is that the girl's performance on the test was ambiguous; perceivers who viewed the testing tape exclusively gave highly variable ratings of her ability. However, after viewing the video, participants who believed she came from a low SES rated her performance as significantly poorer than those who believed she came from a high SES. Additionally, participants assigned to the high SES group reported that the test was more difficult, remembered that the girl had answered more problems correctly, and reported that instances of good performance (rather than poor performance) were more informative as to her actual competencies than did individuals assigned to the

low SES group. In fact, the two groups explicitly framed the same behaviors in different ways: participants reported that the child had "difficulties accepting new information" if she appeared to come from a low SES, but an "ability to apply what she knows to unfamiliar problems" if she appeared to come from a higher SES. That is, they interpreted her ambiguous performance in a manner consistent with stereotypes regarding SES.

To summarize, Darley and Gross (1983) showed that information that *ought* to be irrelevant, and is even *considered to be* irrelevant by the perceiver, dramatically impacts how ambiguous information from a source with higher diagnostic potential (i.e., behavior) is interpreted. As behavior is considered a valid source of information, perceivers will be confident that their judgments are fair, and may be oblivious to the influence of implicit biases. It is possible that facial trustworthiness may exert a similar effect, such that perceivers may temper their negative biases towards an untrustworthy appearing person in the absence of behavioral information, but may interpret ambiguous behavioral information in such a manner as to justify their implicit biases. Thus, people may more negatively evaluate someone who appears untrustworthy when they have information they can use to confirm their prior expectations.

The Present Study

The present study was designed to identify the extent to which facial trustworthiness influences the interpretation of ambiguous behavioral information. To test this, participants were presented with photographs of faces that varied in facial trustworthiness. Each photograph was paired with a vignette, which described a person engaging in a series of ambiguous behaviors. One vignette described a passive person,

whose behavior could not be interpreted as hostile. The other vignette described an assertive, ambiguously hostile person. The goal of the present study was to determine whether the effect of an untrustworthy face would be magnified when paired with assertive, ambiguously hostile behavioral information, which could be interpreted in a manner to confirm participants' negative expectancies regarding an untrustworthy appearing person.

Hypotheses

H1. For both vignettes, participants assigned the trustworthy appearing faces condition will rate the targets as less hostile, less aggressive, and more friendly, kind, considerate, thoughtful, likeable, and trustworthy than participants assigned to the untrustworthy faces condition (i.e., between-subjects, main effect of condition). This would serve as a replication of the effect of facial trustworthiness on interpersonal evaluations observed in previous research (e.g., Porter et al., 2010).

H2. In both trustworthiness conditions, participants will rate targets as more hostile, more aggressive, and less friendly, kind, considerate, thoughtful, likeable, and trustworthy when paired with the assertive vignette than when paired with the passive vignette (i.e., within-subjects, main effect of vignette condition). This would establish that there is greater potential for perceivers to identify negative behavioral information from the assertive ambiguously hostile vignette than the passive vignette.

H3. There will be an interaction between the facial trustworthiness condition and the vignette type, such that the differences between the aggressiveness, hostility, friendliness, kindness, considerateness, thoughtfulness, likeability, and trustworthiness ratings of participants assigned to the untrustworthy faces condition and those of

participants assigned to the trustworthy faces condition will be larger following the assertive ambiguously hostile vignette than following the passive vignette. This was the primary hypothesis of the present study, and would demonstrate that ambiguous behavioral information may intensify evaluations based on facial trustworthiness.

Methods

Participants

The present study was approved by the Institutional Review Board at Oregon State University. Participants (N = 121, 73 male, 45 female, 3 other) were recruited via Prolific (www.prolific.co), a crowdsourcing research platform. Inclusion criteria included normal or corrected to normal vision, fluency in English, and being at least 18 years of age. Participants self-reported eligibility and were compensated with $2.50 for completing the study, which was estimated to take 15 minutes. On average, participants took 12.1 minutes to complete the study (SD = 6.7, min = 4.5, max = 43.7). Participants' mean age was 25 years (SD = 7, min = 18, max = 56). Of the 121 participants, 98 were White, 1 was Black, 18 were Latinx, 5 were Asian or Pacific Islander, 3 were multiracial, and 2 selected "other" (Middle-Eastern, Polish; note, these categories were not mutually exclusive). Of those 121 participants, 9 requested that their data be withdrawn from the study and 5 failed to pass attention checks (note, these were not mutually exclusive).

Request to withdraw data (N = 9) was not significantly correlated with ethnicity or gender. Participants who requested to have their data withdrawn from the study were not significantly older than those who did not, $t(119)=6.15$, p = .77, and they trended toward taking less time to complete the study $t(119)=1.404$, p = .09, MD = 3.12 minutes, SE = 2.22 minutes. Note that in the feedback section of the survey, one participant

indicated that they had erroneously clicked on the "withdraw my data" button and in fact did want their data included, which may have happened to other participants who did not self-report the error. Failing to pass the attention check ($N = 5$) was not associated with age, gender, ethnicity, or time spent on the study.

The final sample consisted of 108 participants (63 men, 42 women, 3 other), with an average age of 25.1 years ($SD = 7.5$, $min = 18$, $max = 56$). Of them, 88 identified as White, 1 as Black, 18 as Latinx, 4 as Asian or Pacific Islander, 3 as Multiracial, 2 as other (Middle Eastern, Polish). Of the final sample, the average time to complete study was 12.2 minutes ($SD = 6.9$, $min = 4.5$, $max = 43.7$).

Measures

Participants' evaluations of the target person in each condition was assessed at two levels: perceptions of the target's personality, and their predicted behavioral intentions toward the target.

Personality Assessment

Participants rated targets on seven personality traits on a Likert a scale from 1 to 7. All traits related to the target's approachability and benevolence, which are key components of evaluations based on facial trustworthiness (Todorov et al., 2008). These included: hostility, aggressiveness, friendliness, likeability, kindness, considerateness, thoughtfulness (Srull & Wyer, 1981), and trustworthiness. Additionally, participants rated the target's intelligence, which was not hypothesized to be associated with evaluations of a targets' trustworthiness, as a check for discriminant validity. Each trait was presented in random order. See Appendix A for the full personality assessment measure.

Behavioral Intent Measure

Participants' behavioral intent toward the target was assessed in a number of hypothetical scenarios, including how likely participants would be to interact with the person described in the vignette if given the opportunity in a social or professional context, how suspicious they would be of the target under various circumstances, and how harshly would they punish the target in a criminal context. The first author (Z.A.) generated 14 items to probe these various dimensions of trustworthiness. Example items include the following:

> *If you were sitting in a waiting room with him, how likely would you be to chat with him?*
>
> *If you were hiring for a job, and this person was a qualified candidate, how likely would you be to hire him?*
>
> *If you wanted to step away from your backpack at an airport, how willing would you be to ask him to watch it for you while you're gone?*
>
> *If you were a judge and this person was found guilty of petty theft (an item less than $100), how much do you think this person should be fined?*

All 14 items were presented to participants in random order. Participants reported their predicted behavioral intent for each item on a 1-7 Likert scale. See Appendix A for the full behavioral intent measure.

Attention Checks

An attention check was included in a random position in the behavioral intent measure (see Appendix A). Since each participant filled out the behavioral intent measure twice (see Procedures), participants had two opportunities to pass or fail the

attention check. Failure to pass either attention check resulted in exclusion from analyses.

Stimuli

Target Faces

Faces were drawn from the Chicago Face Database, an open-access repository of normed facial stimuli (CFD; Ma, Correll, & Wittenbrink, 2015). Two trustworthy and two untrustworthy faces were selected from each of the four ethnic groups included in the Chicago Face Database (i.e., as defined by the CFD, Black, White, Latinx, and Asian), for a total of 16 faces. Note that this study was not designed to parse the effect of ethnicity or interactions between ethnicity and other variables on dependent measures. Rather, since the features associated with facial trustworthiness can vary independently of ethnicity (Birkás et al., 2014; Wilson & Rule, 2015), the goal was to determine the effect of facial trustworthiness on perceivers' evaluation of targets regardless of ethnicity.

Faces were selected based on the normed ratings of facial trustworthiness, age, and attractiveness that accompanied each face in the Chicago Face Database. Facial trustworthiness and attractiveness were rated on a 1-7 scale by raters recruited for the Chicago Face Database. Primary criteria for selection in the present study was a low or high facial trustworthiness rating compared to the other faces in the Chicago Face Database. However, faces were also selected to minimize differences between facial trustworthiness groups in perceived age and attractiveness. That is, the most and least trustworthy faces were selected from each ethnic group, unless a face with a similar facial trustworthiness rating was available with an age or attractiveness rating closer to the group average.

Mean facial trustworthiness was 2.55 (*min* = 2.37, *max* = 2.81) for the eight faces selected in the *untrustworthy* group. Mean facial trustworthiness for the eight faces in the *trustworthy* group was 4.17 (*min* = 3.89, *max* = 4.57). A t-test demonstrated that the two groups were significantly different from one another in the expected direction, *MD* = 1.61, $t(14)=15.372$, $p < .001$.

Average perceived age was 27 years in the *untrustworthy* group, with mean attractiveness at 2.34 (*min* = 1.85, *max* = 3.16). Mean perceived age was 27 years in the *trustworthy* group, and mean attractiveness was 3.70 (*min* = 2.40, *max* = 4.85). Note that although attractiveness was on average higher in the trustworthy group, the range of scores crossed (i.e., some faces in the trustworthy group were rated as less attractive than some faces in the untrustworthy group).

See Figure 12 for an example of a face from each group. See Appendix B for all 16 faces selected. Note that although faces were not selected for physical characteristics (e.g., brow height), all faces in the trustworthy group have either raised brows, upturned lips, or a soft jaw, and all faces in the untrustworthy group have either heavy brows, downturned lips, or a strong jaw (features characteristic of a trustworthy versus untrustworthy appearance; see Figure B).

Figure 12. An example of a face from the trustworthy (left) and untrustworthy (right) groups.

Target Vignettes

Vignettes were based on the classic Donald vignette, which described a person (Donald) engaging in a series of ambiguous, assertive behaviors that could be interpreted as hostile (Srull & Wyer, 1981). Traditionally, this vignette was used to assess priming effects, wherein participants primed with hostile words described Donald as more hostile than those exposed to neutral words. The Donald vignette was used as a template for the stimuli in the present study.

Vignette Development. The first author (Z.A.) wrote three vignettes based on the Donald vignette. All vignettes involved the same social situations (e.g., a request for a donation, a confrontation with a salesperson). One of the three new vignettes described an assertive person encountering social situations similar to the Donald vignette, but in a different order and under slightly different contexts (for example, the target person was asked to donate money rather than blood). The remaining two vignettes matched exactly the original Donald vignette and the new, assertive vignette respectively, except that the target person behaved in passive rather than assertive ways.

Note that the name Donald was replaced with the names Jack, John, James, and Joseph in order to 1) provide names with which to uniquely reference the target of each vignette and 2) avoid the contemporary political connotations of the name "Donald". See Appendix C for all four vignettes.

Vignette Pilot Testing. In a pilot study, $N = 15$ participants (5 female) recruited from the Prolific platform viewed all four vignettes and rated them on the personality assessment measure described above. For each of the seven personality traits, both passive vignettes were rated more positively (e.g., less hostile and more trustworthy) than the assertive vignettes. For all paired t-tests, $t(14)$ ranged from 2.256 to 5.775, the mean difference across passive versus assertive vignettes on a 7 point Likert scale ranged from 1.3 to 2.7, and *p*s ranged from <.001 to .04. As predicted, there was no statistically significant difference in perceived intelligence across vignette condition, $t(14) = 0.155$ to 0.445, $p = .663$ to .879.

Interestingly, the man described in the original Donald vignette ("James") was perceived as less trustworthy ($t(14) = 1.333$, $MD = 1.333$, $SE = 0.287$, $p < .001$), likeable ($t(14) = 3.162$, $MD = 0.667$, $SE = 0.211$, $p = .007$), and considerate ($t(14) = 3.214$, $MD = 0.733$, $SE = 0.228$, $p = .006$) than the man described in the assertive vignette written for this study. This may be because the target person in the original vignette lies about having diabetes in order to avoid donating blood, whereas the target of the new assertive vignette simply refuses to donate.

Given that mean ratings of personality traits were closer to neutral in the new assertive vignette and therefore more ambiguous, the new assertive vignette was selected as a stimulus for the present study over the original Donald vignette. The passive

vignette based directly on the original Donald vignette was selected as the second stimulus vignette.

Study Procedures

Following informed consent, participants in the main study sample were randomly assigned to the trustworthy or untrustworthy condition (see Figure 13). Participants were presented with the assertive and passive vignettes in random order. Each vignette was paired with a face that matched the condition (either trustworthy or untrustworthy), randomly selected from one of the four ethnicity groups. Participants never saw the same ethnicity twice (i.e., if assigned ethnicity A for the first vignette, the participant would see either ethnicity B, C, or D paired with the second vignette). After viewing the photograph and reading the corresponding vignette, participants rated their impressions of the target on the personality assessment and behavioral intent measures.

Thus, targets' facial trustworthiness was manipulated *between-subjects,* vignette type (passive or assertive) was manipulated *within-subjects,* and target's ethnicity was balanced both within and between-subjects.

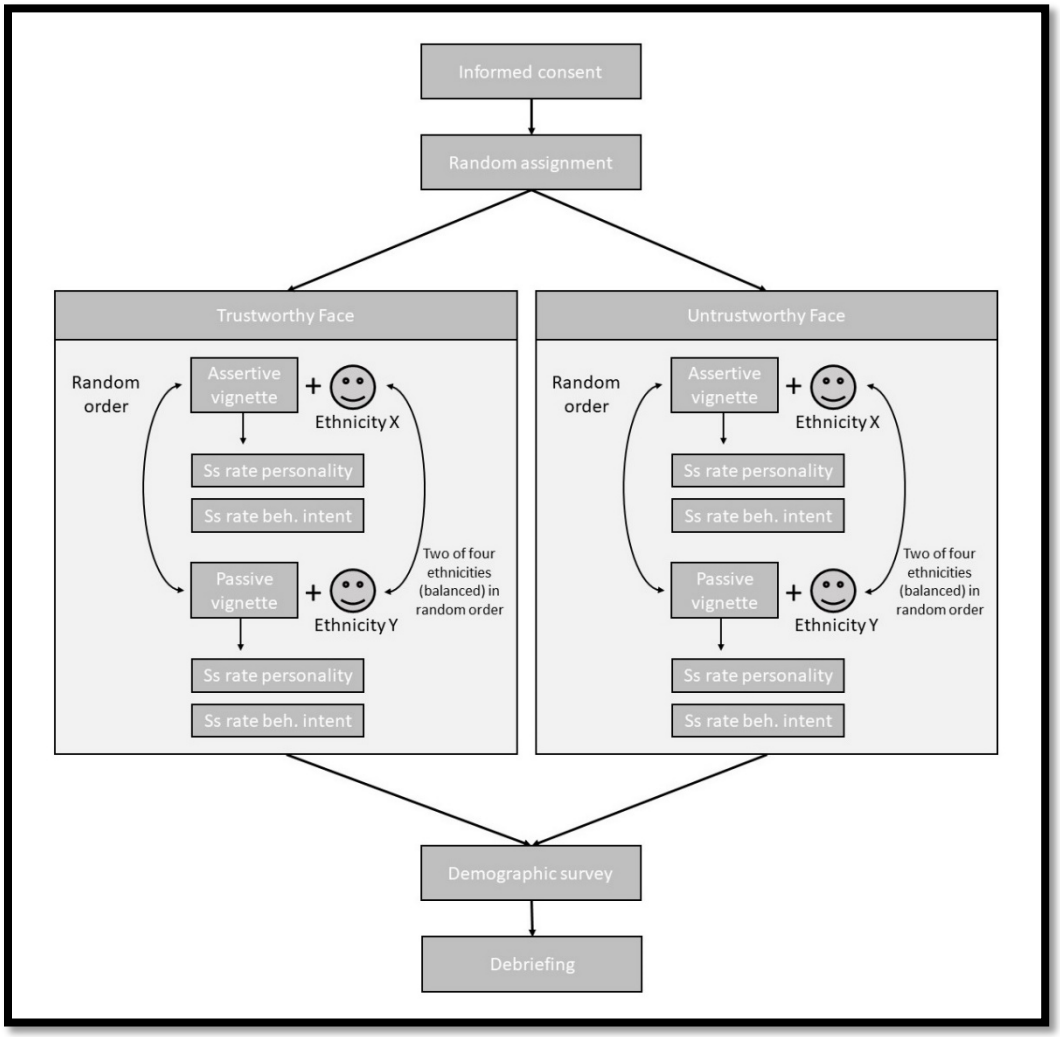


Figure 13. Procedures for Study 2.

Data Analysis Plan

The present study investigated whether perceivers' impressions of a target were influenced by facial trustworthiness when ambiguous assertive versus ambiguous passive behavioral information about the target was provided. Target faces that appeared more trustworthy were hypothesized to be rated as lower on hostility and aggressiveness, and higher on friendliness, kindness, considerateness, thoughtfulness, likeability, and trustworthiness by participants than those who appeared less trustworthy (H1). Furthermore, a main effect of the vignette type was hypothesized, such that targets would

be rated more positively when paired with the passive as opposed to the assertive and ambiguously hostile vignette (H2). Finally, an interaction between facial trustworthiness and behavioral information about the target was predicted, such that participants would rate untrustworthy appearing faces as even more hostile than trustworthy appearing faces when presented with the assertive ambiguously hostile vignette than when presented with the passive vignette (H3).

Hypotheses 1, 2, and 3 were tested using a two-way mixed effects ANOVA. A main effect of *trustworthiness* between-subjects, a main effect of *vignette* within-subjects, and an interaction between both conditions across subjects were predicted for each dependent outcome.

Results

Descriptive Statistics

Personality Assessment

Participants utilized the full 1-7 rating scale for all traits except trustworthiness (*min* = 2, *max* = 7). See Table 7 for descriptive statistics and t-tests for each variable in the personality assessment by vignette type (i.e., passive versus assertive). Consistent with H2, participants rated the target more favorably on all traits except intelligence when the target was paired with the passive vignette.

Table 7. *Descriptive statistics for personality items by vignette condition.*

	Passive		Assertive		Paired T-test		
	Mean	SD	Mean	SD	t	df	p
Hostility	1.79	1.19	4.19	1.40	-13.938	106	<.001
Trustworthiness	4.69	1.18	3.88	1.08	5.421	106	<.001
Friendliness	5.33	1.07	3.18	1.16	14.20	106	<.001
Likeability	5.01	1.23	3.19	1.22	10.24	106	<.001
Considerateness	4.98	1.42	3.03	1.42	9.19	107	<.001
Kind	5.26	1.17	2.91	1.15	13.78	107	<.001
Aggressive	1.58	0.88	4.05	1.40	-15.36	107	<.001
Thoughtful	4.49	1.28	3.28	1.16	6.00	106	<.001
Intelligence	4.11	1.13	4.26	1.13	-0.96	107	0.338

Behavioral Intent Measure.

To allow for items to be consolidated across the behavioral intent measure, negatively worded items (e.g., *If you saw this person standing around in your neighborhood for long periods of time, how likely would you be to report him as a suspicious person?*) were reverse coded. Although internal consistency across the 14 behavioral intent items was strong (Cronbach's $\alpha = .83$ in the assertive condition; $\alpha = .78$ in the passive condition), a principle components analysis with Varimax rotation indicated there was an underlying factor structure to the measure. In both the assertive and passive conditions, four factors emerged with Eigen values > 1, explaining 70% and 65% of item variance, respectively. These factors mapped on to an interpersonal responsibility dimension (five items; i.e., would trust the person to fulfil a social role, such as returning lent money or performing well at a new job), a social engagement dimension (three items; i.e., would engage with this person socially), a suspicion dimension (three items; i.e., would suspect this person of committing an infraction), and a punishment dimension (three items; i.e., would punish this person harshly for

infractions). See Appendix D for factor loadings across components and the final factor structure.

Because of these factor loadings and inter-item correlations, the 14 items of the behavioral intent measure were consolidated into four variables, each a mean of the items indicated on each factor: behavioral intent (BI) responsibility, BI social, BI suspicion, BI punishment. Participants reported a more positive behavioral intent toward the target when the target was paired with the passive vignette for each of the behavioral intent dimensions; see Table 8 for descriptive statistics and t-tests of these four items.

Table 8. *Descriptive statistics for behavioral intent items by vignette condition*.

	Passive		Assertive		Paired T-test		
	Mean	SD	Mean	SD	t	df	*p*
Social	4.49	1.43	3.24	1.50	8.222	107	<.001
Responsibility	4.62	1.13	3.39	1.42	7.693	107	<.001
Suspicion	4.99	1.07	4.31	1.22	6.794	107	<.001
Punishment	5.54	1.00	5.06	1.15	5.792	107	<.001

A Priori **Hypothesis Tests**

Participants in the trustworthy faces condition were hypothesized to evaluate targets more favorably than those in in the untrustworthy faces condition (between-subjects; H1). Vignette condition was also hypothesized to predict participants' assessments of targets, such that participants would evaluate targets more favorably in the passive condition than the assertive condition (within-subjects; H2). Finally, facial trustworthiness condition was hypothesized to interact with vignette condition, such that the effect of an untrustworthy appearance would be most influential when participants had ambiguous behavioral information about the target (i.e., the assertive vignette; H3).

Contrary to study hypotheses and prior literature (e.g., Porter et al., 2010; Wu et al., 2018) there was no main effect of facial trustworthiness on the outcome variable of participants' ratings of the target's trustworthiness, $F(1,105) = 0.133$, $p = .716$. That is, participants' evaluations of the trustworthiness of targets who were paired with the faces with the highest and lowest trustworthiness ratings from the Chicago Face Database normed dataset were not different from one another. Consistent with study hypotheses, there was a main effect of vignette type, such that participants evaluated targets as more trustworthy when paired with the passive vignette, $F(1,105) = 28.822$, $p < .001$. However, contrary to study hypotheses, there was no interaction between facial trustworthiness condition and vignette type, $F(1,105) = 0.818$, $p = .368$.

Although participants evaluated targets more favorably in the passive (as opposed to the assertive) vignette condition in each case, with the single exception of ratings of intelligence, a main effect of trustworthiness was not observed in any personality or behavioral intent outcome (see Table 9). There was no evidence that participants utilized target's facial trustworthiness in their evaluations of the target for any outcome, and there was no evidence that targets' facial trustworthiness interacted with the vignette condition.

Table 9. *Tests for a main effect (between-subjects) of facial trustworthiness condition.*

Outcome	F	df H	df Error	*p*
Trustworthiness	0.133	1	105	0.716
Hostility	2.943	1	105	0.089
Aggressiveness	0.731	1	106	0.394
Kindness	1.250	1	106	0.266
Considerateness	1.838	1	106	0.178
Thoughtfulness	1.250	1	106	0.266
Likeability	1.291	1	105	0.258
Friendliness	2.265	1	105	0.135
BI Social	0.035	1	106	0.851
BI Competence	0.858	1	106	0.356
BI Suspicious	0.149	1	106	0.700
BI Punish	0.519	1	106	0.473

Post Hoc **Analyses**

The present study failed to replicate the effect of facial trustworthiness on participants' evaluations of a target, an effect which has been demonstrated in numerous person perception studies (e.g., T. Li et al., 2017; Porter et al., 2010; Shen et al., 2020; Wilson & Rule, 2015; Wu et al., 2018). This failure to replicate an established effect precluded conclusions regarding the main study hypothesis (i.e., that ambiguous behavioral information would strengthen the effect of facial trustworthiness).

Because this outcome was unexpected, *post hoc* analyses were conducted to probe for issues in the study design, with the goal of determining what factors may have prevented successful replication of this established effect. These included: reexamination of the target facial stimuli, controlling for target's ethnicity, an analysis of the subset of evaluations made first (i.e., participants' first but not second ratings), and finally, a nested

hierarchical regression model with facial trustworthiness treated as a continuous rather than dichotomous variable.

Test for Idiosyncratic Facial Stimuli

Participants may have reacted to some of the facial stimuli idiosyncratically. That is, factors unique to one or more of the stimulus faces may have introduced error to the study design that masked an effect of facial trustworthiness condition.

Therefore, a graph was created that displayed participants' evaluations of the target by vignette type (Y axis) and the faces' original Chicago Face Database normed facial trustworthiness score (X axis). Graphs for the outcomes of trustworthiness and hostility (reverse coded) are displayed below (see Figures 14 and 15). Each blue dot represents a separate face in the passive vignette condition, with a corresponding orange dot representing the same face in the assertive vignette condition. Note that the faces clustered between values 2.1 and 3.1 represent faces in the untrustworthy group, and faces clustered between 3.6 and 4.6 represent faces in the trustworthy group.

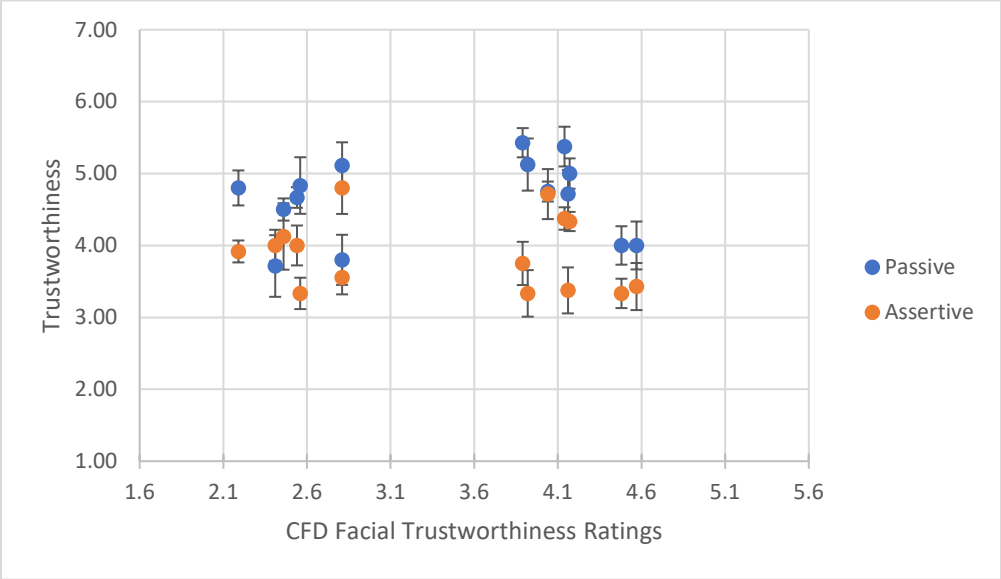


Figure 14. Face stimuli plotted by participants' assessment of their trustworthiness and Chicago Face Database facial trustworthiness ratings.

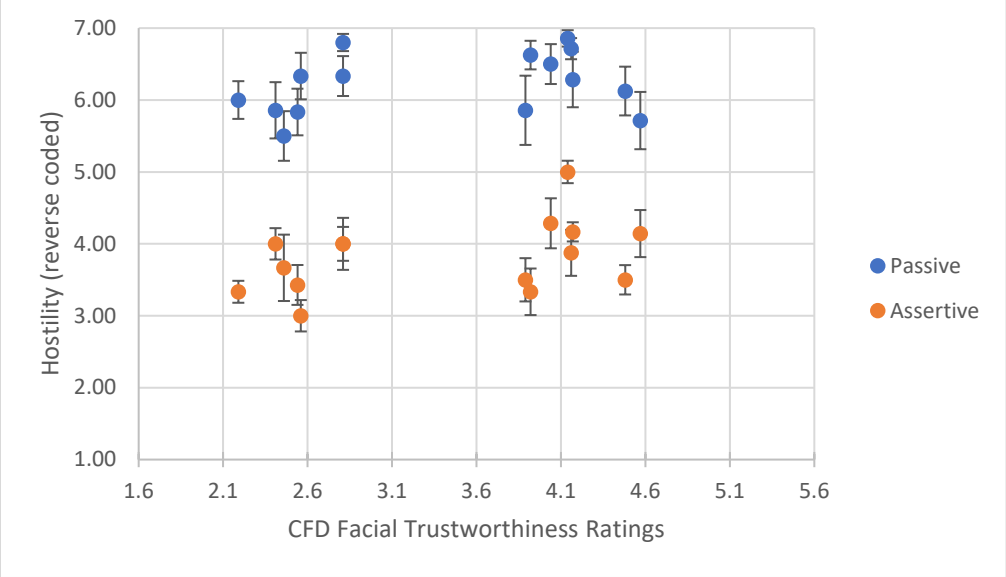


Figure 15. Face stimuli plotted by participants' assessment of their hostility (reverse coded) and Chicago Face Database facial trustworthiness ratings.

Although there was variation across faces in both participants' evaluations and the Chicago Face Database trustworthiness score, no faces demonstrated an especially strong effect of vignette or an especially aberrant outcome score. Furthermore, there was no

positive linear trend, whereby participants' evaluations of the target become more positive as Chicago Face Database trustworthiness ratings increased. Thus, these graphs provided more evidence that participants in this study were not sensitive to targets' facial trustworthiness and indicated that idiosyncratic characteristics of a particular face did not drive the null effects observed.

Controlling for Ethnicity

Few lab studies investigating the effect of facial trustworthiness have varied the ethnicity of the target (e.g., Todorov et al., 2009; but see Birkás et al., 2014). Although prior research has shown that the physiognomic characteristics associated with facial trustworthiness can vary independently of ethnicity, some evidence indicates that perceivers are more sensitive to facial trustworthiness when the target's ethnicity differs from their own (Birkás et al., 2014). Although ethnicity was balanced between conditions in the present study, it is possible that including multiple ethnic groups introduced error that masked an effect of facial trustworthiness.

Two-way mixed effects ANCOVAs were run for the two primary personality evaluation items (trustworthiness and hostility), with ethnicity included as a covariate. Facial trustworthiness condition was not a significant predictor of participants' assessment of targets' trustworthiness, $F(1,99) = .027$, $p = .869$. However, Asian ethnicity was a significant predictor of participants' trustworthiness ratings, $F(1,99) = 6.021$, $p = .033$. Facial trustworthiness condition was also not associated with participants' evaluations of targets' hostility, $F(1,99) = 3.609$, $p = .148$. However, Black ethnicity was associated with participants' evaluations of targets' hostility, $F(1,99) = 2.224$, $p = .008$.

These analyses demonstrated that the present study was capable of detecting a between-subjects effect related to characteristics of the facial stimuli (i.e., ethnicity). However, variations in targets' ethnicities were not responsible for masking an effect of facial trustworthiness on participants' evaluations of targets.

Testing Between-Subjects Effects Only

Although the behavioral information in each vignette was designed to be ambiguous, it is possible that the presentation of two vignettes with very different protagonists (i.e., assertive versus passive) would, when contrasted with one another, erode this ambiguity. That is, although the target of the vignette's behavior might at first appear ambiguous, when contrasted with the behavior of the target described in the second vignette, that behavior may become clearly passive (or assertive) to the perceiver. If the information presented is not ambiguous, the role of facial trustworthiness in its interpretation may be less salient, as was found by Shen and colleagues' (2020).

Thus, the subset of data recording perceivers' *first* assessment of a target was analyzed via a two-way between-subjects ANOVA, with both vignette and trustworthiness condition treated as between-subjects variables. Vignettes were presented in random order. Therefore, approximately half (N = 55 out of 108) of participants read the passive vignette first, and these participants comprised the passive condition for this analysis. Again, there was no significant effect of facial trustworthiness condition on any of the study outcomes, $F(1,102) = 0.022$ to 2.666, ps = .106 to .883. These analyses were repeated as an ANCOVA controlling for ethnicity, which produced a similar pattern of results. Thus, there was no evidence of a between-subjects effect of facial trustworthiness on perceivers' evaluations of targets.

Facial Trustworthiness as a Continuous Predictor

Dichotomizing a continuous predictor can result in significant reductions in study power (Altman & Royston, 2006). In the present study, facial trustworthiness was treated as a dichotomous predictor; i.e., faces were categorized as either trustworthy or untrustworthy. Reducing the facial stimuli into two conditions for analyses ignored variance associated with individual faces. Indeed, the faces selected for analysis varied along a continuum of facial trustworthiness, as indicated by normed data from the Chicago Face Database (see Figures 14 and 15). Therefore, a hierarchical linear regression (nesting observations within participants to control for nonindependence) was run for each personality assessment and behavioral intent variable.

Vignette type and Chicago Face Database ratings of attractiveness (a potential confound) were input as covariates in the model, in addition to the independent variable of Chicago Face Database facial trustworthiness ratings. The models were run both with and without an interaction term for vignette and Chicago Face Database facial trustworthiness.

Statistical tests for two outcomes resulted in a p value close to .05, but only in the models without an interaction term for facial trustworthiness condition and vignette. Facial trustworthiness was associated with ratings of hostility, $\beta = 0.34$, $SE = .17$, $p = .047$, and BI Social, $\beta = -.35$, $SE = .19$, $p = .066$, at borderline significance. Betas for all other outcomes ranged from |.002| to |0.22|, ps = .181 to .990. Given that there was no consistent effect of facial trustworthiness across outcomes, it was concluded that the limited sensitivity of a binary versus continuous predictor of study outcomes was not responsible for the null effect observed in the present study.

Discussion

The present study used an experimental paradigm to investigate the influence of target's facial trustworthiness when the perceiver is presented with ambiguous behavioral information about the target. Failure to replicate the established effect of facial trustworthiness on perceiver evaluations of the target precluded analysis of the primary hypothesis (i.e., that ambiguous behavioral information would amplify the effect of facial trustworthiness on perceiver evaluations of the target). *Post hoc* analyses did not provide evidence that this failed replication was due to target ethnicity, presentation order, idiosyncrasies of the facial stimuli, or dichotomizing facial trustworthiness.

Study Strengths

Despite these null effects, the present study was characterized by several strengths. Pilot testing ensured that the assertive vignette was indeed perceived ambiguously in terms of trustworthiness and hostility by perceivers, and statistical tests within the primary study indicated that the vignette successfully manipulated participants' evaluations of the targets along a *benevolence* dimension (i.e., trustworthiness, hostility, aggressiveness, kindness, likeability, etc.) but not *competence* (i.e., intelligence). Facial stimuli from the Chicago Face Database were highly standardized, such that camera angle, photograph backgrounds, and facial expression were the same across faces. Normed data accessed through the Chicago Face Database ensured that the faces selected as stimuli were perceived of as highly trustworthy or highly untrustworthy, and that across groups, differences in perceived age and attractiveness could be minimized (Ma et al., 2015). Additionally, the utilization of real human faces (as opposed to computer generated models, as leveraged for increased

experimental control by researchers such as Shen et al., 2020, and Todorov et al., 2008) improved the study's external validity.

Limitations

That a main effect of facial trustworthiness was not observed is theoretically surprising and inconsistent with the literature. Examination of the aspects of study design unique to this project, as compared to studies that successfully demonstrated an effect of facial trustworthiness on perceivers' evaluations, may illuminate the circumstances underlying the observation of a null effect.

First, most studies that successfully detected an effect of facial trustworthiness presented photographs of targets from a single racial group, most commonly Caucasian (e.g., Jaeger et al., 2019; Porter et al., 2010; van 't Wout & Sanfey, 2008), Asian (e.g., T. Li et al., 2017; Wu et al., 2018), or digital models reminiscent of Caucasians (e.g., Jaeger et al., 2019; Shen et al., 2020; Todorov et al., 2008, 2009). Ethnicity was balanced across conditions in the present study, allowing for generalizability of study findings across multiple groups. However, it is possible that the introduction of ethnic variation gave participants an erroneous interpretation of the purpose of the study, or simply introduced another source of variance that may have affected the detection of an effect of facial trustworthiness. Indeed, there is some evidence that perceivers are more sensitive to facial cues of trustworthiness in the faces of targets of an ethnic group different from their own (Birkás et al., 2014). Yet it is possible to detect an effect of facial trustworthiness when targets are of mixed ethnicity. For example, Wilson and Rule (2015) collected facial trustworthiness ratings of both Black and White male convicts, and determined that 1) facial trustworthiness was associated with their sentencing

outcomes and 2) the relationship between sentence and facial trustworthiness was not moderated by target race in that sample (Wilson & Rule, 2015).

In the present study, facial trustworthiness was manipulated between-subjects. This design was chosen so that ethnicity could be varied within-subjects without confounding facial trustworthiness condition. Typically, the effect of facial trustworthiness is consistent, but small (Todorov et al., 2015), and a within-subjects design is better suited to detect small effects (Greenwald, 1976). Indeed, it was a within-subjects manipulation of targets' facial trustworthiness that resulted in the detection of statistically significant differences in investment (Jaeger et al., 2019; T. Li et al., 2017; van 't Wout & Sanfey, 2008) and sentencing decisions (Porter et al., 2010; Wu et al., 2018) in a laboratory context. Failure to detect a main effect of facial trustworthiness may be partially attributable to this study's between-subjects design.

Additionally, vignette type (assertive versus passive) was manipulated within-subjects, such that each participant read both vignettes. Key to the present study's hypotheses was the ambiguous nature of those vignettes. However, presenting both vignettes to participants allowed them to serve as contrasts to one another, eliminating the potential for participants to interpret them ambiguously. Indeed, Porter and colleagues (2010) manipulated vignette type between-subjects (although the contents of that vignette described a crime that varied in severity, rather than ambiguous behavior that varied in assertiveness) and facial trustworthiness within-subjects, opposite design decisions than those made for the present study.

Future Research

A more effective study design to tackle the primary hypothesis of the present study may follow the lead of Porter and colleagues (2010), such that facial trustworthiness condition is manipulated within-subjects and vignette type is manipulated between subjects. Ideally, four vignettes would be created, two passive and two assertive vignettes that are rated similarly by perceivers in pilot work. Ethnicity of targets would be manipulated between-subjects, such that each participant views two faces from the same ethnic group that vary in facial trustworthiness. Although a descriptive vignette is clearly different from the sort of behavioral information gleaned from real interactions with real people in day to day life, such a future study is a first step to illuminating the extent to which ambiguous behavioral information amplifies the effect of facial trustworthiness on perceivers' evaluations of the target of their perceptions.

General Discussion

The field of person perception has demonstrated that appearance-based judgments are rapid, implicit, and inevitable, yet only marginally associated with the behavioral tendencies of the person perceived (Todorov, 2015). In this manner, people who look untrustworthy are treated with more suspicion and punished more harshly than those who appear trustworthy, both in the lab (e.g., Wu et al., 2018) and *in vivo* (e.g., Wilson & Rule, 2015, 2016). Yet surprisingly little research has investigated how an untrustworthy appearance might be associated with a person's outcomes throughout the lifespan, or how a relationship between appearance and behavior might develop over time. Such research may have implications for other fields of psychology in addition to person perception, such as developmental psychopathology.

Zebrowitz' model (see Figure 1) provides a framework through which the various mechanisms linking facial appearance and outcomes for the target might be assessed. For example, an untrustworthy appearing person may elicit negative behavioral responses from social partners, causing the target to behave negatively and confirm the perceivers' negative expectancies (expectancy effects). Additionally, engagement in delinquent behavior might itself have deleterious effects on the face, such that over time a person begins to appear less trustworthy (Dorian Gray effects). To investigate each link in Zebrowitz' model, it is necessary to investigate both perceivers and targets. That is, it is necessary to study the processes by which an untrustworthy appearance develops within targets, and how such an appearance may expand or limit the target's behavioral opportunities. Furthermore, the factors contributing to perceivers' appearance-based judgments, how those judgments influence their behavior toward a target, and how the

target's appearance affects perceivers' interpretations of other information about the target must each be analyzed.

Study 1 leveraged methods from developmental psychopathology to investigate the target by assessing facial trustworthiness of 206 at-risk youth and following the development of their facial trustworthiness and engagement in delinquency across two decades. Thus, Study 1 measured targets' appearance and behavior in search of patterns consistent with Dorian Gray effects (Pathway C; see Figure 1) and expectancy effects (Pathways d & D; see Figure 1). Study 2 used experimental psychology methods common in the person perception literature to investigate the processes by which appearance might influence the social environment (Pathway d; see Figure 1), i.e., how a target's appearance might influence perceivers' interpretation of ambiguous behavioral information about the target. Both studies present clear directions for future research.

Future Research

To advance both the theoretical understanding and potential applications of Dorian Gray and expectancy effects related to facial trustworthiness, it is necessary to: 1) identify the mechanisms theoretically underlying these effects, 2) examine the key social contexts in which relations between facial trustworthiness and delinquency may unfold, 3) apply this model to other populations than that observed in the present project, and 4) collect longitudinal data on targets' appearances.

Identifying Theoretical Mechanisms of Zebrowitz' Model

Study 1 demonstrated that facial trustworthiness during early adolescence was associated with rate of change in delinquency across development, and that furthermore, early levels of delinquency can predict changes in facial trustworthiness. However, the

microsocial interactions thought to underlie expectancy effects and the mechanisms of change in facial appearance underlying Dorian Gray effects were not measured. Thus, the iterative processes that contribute to these effects over time are as of yet unknown. The present project provided evidence consistent with two of the theoretical pathways identified in Zebrowitz' model (see Figure 1), creating an impetus to examine the intermediary processes that may causally link facial trustworthiness and uncooperative or "untrustworthy" behavior. The examination of these theoretical processes may be applicable across many populations and settings beyond that studied in the present project.

Mechanisms of Expectancy Effects. A crucial next step for future research regarding facial trustworthiness and expectancy effects is to investigate how facial trustworthiness might impact the social environment (Pathway d in Zebrowitz' model; see Figure 1). Laboratory research could close this gap by determining the extent to which facial trustworthiness influences perceivers' interpretations of ambiguous target behaviors, or by utilizing Gunaydin and colleagues' (2017) paradigm, in which participants' judgments of another person's likeability based on a facial photograph predicted how they approached the person one month later and how much they reported actually liking that person after meeting. Rather than likeability, participants could rate the target's trustworthiness.

Future research should also examine how a perceiver's face-based target evaluations and subsequent behavior might influence the behavior of the target (Pathway D; see Figure 1). This could be done by modifying Snyder and colleague's (1977) paradigm, which found that men spoke more warmly to and elicited more warm

responses from women they were led to believe were attractive than women they were led to believe were unattractive, regardless of the women's actual appearance. In this case, the attractiveness manipulation would be replaced with a facial trustworthiness manipulation. Furthermore, the paradigm could be extended to determine whether an interaction experienced by the target with a perceiver who believed them to appear untrustworthy might alter the target's future behavior—for example, their likelihood to cheat a partner in a lab-based task.

There are ethical and practical issues preventing the elicitation of delinquency in a laboratory environment. However, longitudinal research could connect observations of expectancy effects related to facial trustworthiness to delinquency tangentially. For example, targets' reports of negative social interactions (such as conflictual interactions with law enforcement officers, teachers, or cashiers) across the lifespan could be examined in relation to their facial trustworthiness and self-reported delinquency. Targets' perceptions of these interpersonal interactions may be skewed, yet in conjunction with the proposed research agenda outlined above, they provide valuable insight into how the target's social experiences may be associated with facial trustworthiness.

Each of these examples would advance a theoretical understanding of the microsocial processes that may underlie the longitudinal effects observed in Study 1.

Mechanisms of Dorian Gray Effects. In Study 1, youth with higher levels of delinquency at early adolescence experienced more rapid and pronounced decrements in facial trustworthiness than their peers who had engaged in less delinquency. It remains unclear what factors drove the accelerated decreases in facial trustworthiness observed in

delinquent youth. However, it is notable that decrements in facial trustworthiness were observed across adolescence and not adulthood, indicating that adolescence may be a formative developmental period for facial trustworthiness. Thus, the investigation of Dorian Gray effects may benefit from closely examining this developmental timepoint. Future research should isolate the causal mechanisms that may underlie degradations in facial trustworthiness as they relate to delinquency during adolescence.

Although the complexity of the analytical models in Study 1 precluded the analysis of substance use alongside the parallel processes of delinquency and facial trustworthiness, Alley and colleagues (2019) found that tobacco use (which covaried with delinquency) predicted decrements in facial trustworthiness by early adulthood in this sample. Indeed, smoking is associated with changes in the wrinkling of the lips and eyes that may affect perceived facial trustworthiness (Okada et al., 2013). It is possible that tobacco use was more frequent among youth who engaged in delinquency, such that as youth took part in delinquent behaviors and formed delinquent peer groups, they began to smoke more, and that smoking led to decrements in facial trustworthiness. It may also be that the relationship between delinquency and facial trustworthiness is spurious, and only a marker of these variables' shared association with tobacco use.

To parse these possibilities, substance use across time should be examined with more granular detail than that in Alley and colleague's (2019) study. Tobacco use at several developmental timepoints during early adolescence could be examined as a predictor of rate of change in facial trustworthiness. Future research should also examine whether early initiation in tobacco use, frequency of tobacco use, or level of tobacco use (i.e., whether smoking when young, smoking often, or smoking a lot) better predict

decrements in facial trustworthiness. Finally, researchers could examine whether association with deviant peers mediates the associations between delinquency, tobacco use, and facial trustworthiness.

It is also possible that frequent facial expressions may drive change in appearance by altering the wrinkling and musculature of the face (Malatesta et al., 1987). Adolescents who engage in delinquency and associate with deviant peers may express more negative facial expressions, which could culminate in a face that at rest resembles a scowl—consistent with a prototypically untrustworthy appearance (Todorov et al., 2008).

Given that the Dorian Gray effect is likely to be subtle and to unfold over long periods of time, it is difficult to isolate this process in a laboratory setting. However, future work could film conversations between adolescents involved in the justice system versus those who are not and code facial expressions during these conversations. It may be that justice-involved youth make more negative facial expressions than prosocial adolescents.

Additionally, future work could leverage existing data from successful intervention studies that aim to reduce delinquency. Should these studies include photographs of participants, researchers could test whether facial trustworthiness increased or decreased over the course of the intervention across intervention groups, and whether these changes were mediated by delinquency. Note that as Study 1 indicates that facial trustworthiness declines across adolescence, a comparison group would be necessary in order to determine whether any observed decreases in facial trustworthiness are abnormally rapid or decelerated.

Key Social Contexts for Applications of Zebrowitz' Model

As was observed in Study 1, developmentally, delinquency increases during adolescence (Patterson, 1993; Patterson et al., 1989) and decreases across early adulthood (Sampson & Laub, 2003; Wiesner et al., 2005). Therefore, to reduce delinquency to the benefit of both the actor and those affected by their delinquent behavior, initiation and escalation in delinquency during adolescence must be reduced, and reductions across early adulthood must be increased.

Facial trustworthiness was associated with rate of change in both escalation and degradation of delinquency in this sample, in a manner consistent with expectancy effects. It is possible that the Dorian Gray effect may also impact the development of delinquency, via expectancy effects. That is, low facial trustworthiness early in life may increase engagement in delinquency during adolescence (expectancy effects), leading to a less trustworthy appearance by adulthood (Dorian Gray effect), which may reduce the speed with which delinquency declines across this developmental period (expectancy effects). Thus, expectancy effects and Dorian Gray effects may operate simultaneously, sequentially, and reciprocally, in a manner that impacts the development of delinquency.

Although Study 1 only measured the appearances and behaviors of the targets, given that these are social processes, expectancy effects must manifest through social interactions with perceivers. Indeed, as discussed above, social processes may even mediate Dorian Gray effects (e.g., association with a particular peer group may encourage grooming behaviors, facial expressions, or substance use that lead to an untrustworthy appearance). Therefore, research regarding the development of Dorian Gray and expectancy effects regarding facial trustworthiness and delinquency must focus

on important social partners who are affected by or affect the development of facial trustworthiness and delinquency—for example, teachers and peers.

Teachers. Teachers are important authority figures that have the potential to dramatically shape youth outcomes. Prior research has demonstrated that manipulating teachers' impressions of their students' potential to excel impacts students' actual scholastic achievement (e.g., Rosenthal, 1987). Furthermore, teachers may even influence the development of their students' problem behavior. For example, conflictual teacher-student relationships are associated with the student's later problem behaviors (Rudasill et al., 2010), and critically, in a sample of at-risk youth ages 13-18, positive student-teacher relationships predicted reduced misconduct (Wang et al., 2013).

There is evidence that teachers are influenced by their students' appearances. For example, expectancies regarding students' future scholastic performance are associated with the students' appearances (Dare, 1992). Furthermore, teachers interpret objective scholastic information more favorably when students look attractive (Clifford & Walster, 1973). Teachers' nominations of high performing students are associated with the students' attractiveness (Babad et al., 1982), and finally, physical characteristics of the student impact student-teacher interactions during the first week of school (Adams & Cohen, 1974). Thus, it is possible that students' facial trustworthiness may impact teachers' impressions of students in a manner that could elicit expectancy effects regarding delinquency and other problem behavior.

To determine whether facial appearance is associated with the experiences of adolescents who vary in facial trustworthiness in the classroom, future research could focus on trustworthy and untrustworthy appearing adolescents and gather their report of

perceived student-teacher relationships and frequency of the school's disciplinary action toward them. Additionally, future research could focus on teachers as perceivers. Ambiguous information about a student could be accompanied with a picture of a trustworthy or untrustworthy appearing adolescent and presented to teachers for interpretation. For example, if a student leans over a desk, a teacher could interpret that behavior as either stretching or peeking at a classmate's work. Laboratory research could investigate to what extent students' facial trustworthiness impacts teachers' interpretations of such ambiguous behaviors.

Peers. As association with deviant peers is influential for the development of delinquency (Dishion, 2000; Jessor, 1991; Van Ryzin & Dishion, 2014), future research should investigate the relationship between facial trustworthiness and the formation of peer groups during adolescence. Such an investigation would benefit from a two-pronged approach that tackles both perceivers and targets. Longitudinal research that follows targets across time could determine the extent to which facial trustworthiness predicts their association with delinquent peers. When investigating youth as perceivers, one might examine how facial trustworthiness is interpreted by justice-involved youth. There is evidence that the manner in which perceivers interpret others' physiognomic traits may be shaped by their goals for social interaction (Zebrowitz & Montepare, 2006). Thus, an untrustworthy appearing person may be off-putting to prosocial peers but attractive to delinquent peers, for whom a proclivity to break social norms would be desirable.

Similar studies may be adapted to investigate the role of interactions with other important social figures (e.g., employers and police officers), who could have some

influence over the development or desistence of delinquency, and may themselves be susceptible to the influence of targets' facial trustworthiness. Regardless of the precise social context, expectancy effects and Dorian Gray effects involve reciprocal, social processes. It is therefore paramount to investigate the experiences of both the target and the perceiver.

Applying Zebrowitz' Model to Other Populations

Study 1 examined Dorian Gray and expectancy effects among a sample of at-risk, low income, mostly white (90%) males from ages 13 to 38. It is possible, and perhaps even likely, that the trends observed in this sample may not generalize to other groups. For example, there is some evidence that cooperative or trustworthy behavior is associated with men's but not women's appearances (e.g., Foo et al., 2019), such the effects observed in Study 1 may be inconsequential for the development of delinquency among females. It is critical that future research address the extent to which these effects replicate among populations other than that observed in the present project.

Target Sample and Ethnicity. Perhaps the most obvious direction for future research is to expand the ethnic diversity of the targets. Ethnicity has a powerful influence on person perception, as evident in both prior research (e.g., King & Light, 2019) and the between-subjects effects observed in Study 2. Limited research has investigated the relationship between ethnicity and facial trustworthiness, but the literature that does exist indicates that this relationship is complex. Stanley and colleagues (2011) found that implicit racial bias predicted how trustworthy perceivers rated faces of various ethnicities, and furthermore, how much perceivers were willing to trust partners in a lab task. Birkás and colleagues (2014) found that perceivers were more

sensitive to variations in facial trustworthiness when the ethnicity of the target differed from the ethnicity of the perceiver. On the other hand, Wilson and Rule (2015) found no interaction between target's race and perceived facial trustworthiness, and no moderating effect of race on the association between sentencing outcomes and facial trustworthiness. Future research should examine 1) whether the effects observed in Study 1 are replicable in more diverse samples and 2) the extent to which ethnicity may moderate these associations.

It should be noted that the composition of the target sample has implications for the composition of the perceivers that surround and judge these targets. For example, the Study 1 sample was racially homogenous largely due to the demographic composition of the region from which participants were recruited. Therefore, the perceivers that participants interacted with on a day to day basis were likely of a similar demographic composition. Thus, sampling from a region with a more diverse population would result in both a more diverse population of targets and a more diverse population of perceivers. This distinction is important, as characteristics of the targets and perceivers can interact to affect judgments of facial trustworthiness in two important ways: facial typicality (i.e., similarity of the target's face to faces with which the perceiver is familiar) and the social goals of the perceiver.

Target Sample and Facial Typicality. Although the emotion overgeneralization phenomenon appears to underlie a universal cue to facial trustworthiness, trustworthiness judgments are also sensitive to facial typicality; i.e., how "normal" (or abnormal) a face appears to a perceiver (Todorov et al., 2015; Wilson & Rule, 2015). Perceivers' perceptions of typicality have been manipulated in a laboratory environment, such that

once participants had become accustomed to faces of type "A", presenting faces of type "B" reduced perceivers' trustworthiness ratings, even though this manipulation was orthogonal to emotion-overgeneralization effect cues of facial trustworthiness (Todorov et al., 2015). Thus, in different target samples, there may be different baseline levels of facial trustworthiness or unique appearance cues to trustworthiness related to typicality of faces in that population. Together, these factors may lead to different developmental trends than those observed in Study 1.

Target Sample and Social Goals of the Perceiver. There is evidence that the manner in which perceivers interpret others' physiognomic traits may be shaped by their social goals (Zebrowitz & Montepare, 2006). For example, low facial trustworthiness may be an off-putting appearance characteristic for prosocial perceivers, but attractive to perceivers who are seeking a "partner in crime." The fact that the perceiver's identity and social goals may be important for their interpretation of the target is another reason to replicate Study 1 in more diverse groups of targets, who may be surrounded by different types of perceivers (with different social goals than those of other populations) in addition to being perceived differently (due to different standards of facial typicality).

The Value of Longitudinal Designs to Person Perception Research

In addition to requiring specialized expertise, longitudinal research is time intensive and expensive. These methods are often outside of the purview of person perception researchers, yet the implications of person perception research (as evidenced by this project's findings and Zebrowitz' model) extend into developmental contexts. These theoretical models cannot be tested or applied to disciplines that could benefit from their insight (e.g., developmental psychopathology) without longitudinal research.

Pictures of participants across development are required to test person perception models that imply change across time; indeed, Study 1 was only possible because the researchers responsible for the Oregon Youth Study had the foresight to include photographs of participants across development. Therefore, it would benefit both the field of person perception and other disciplines if longitudinal research, when possible, included pictures of participants. Ideally, these photographs would be taken across development, but even retrospective collection of facial pictures at the end of a study would be beneficial. These photographs could be the foundation for tests of Zebrowitz' model across many different facial features, from attractiveness, babyfacedness, facial trustworthiness, and beyond, that may have implications for many developmental processes.

Summary

Research regarding Dorian Gray and expectancy effects related to facial trustworthiness has the potential to deepen the scientific understanding of the development of delinquency. It may be most important to examine these processes during adolescence, when facial trustworthiness appears to be most malleable. Such efforts may lead to improved prevention and intervention programs.

Future research should prioritize the identification of the microsocial interactions that culminate in expectancy effects, both in the lab and through targets' longitudinal reports. Future research must also examine the influence of tobacco use and deviant peer groups on the development of facial trustworthiness, and furthermore, the role of facial trustworthiness in the formation of deviant peer groups and initiation in substance use during early adolescence. Future research must specifically target important others, such

as peers, teachers, coworkers, employers, and other authority figures that have the potential to impact the development of delinquency through expectancy effects. Furthermore, future research must determine the extent to which these processes apply across demographic and regional contexts. Finally, because of the difficulty of studying delinquency and isolating longitudinal phenomena in a laboratory context, future research will depend upon the inclusion of facial photographs in longitudinal research regarding delinquency.

Conclusion

Person perception doesn't just provide the perceiver with a snapshot of the social world; it actively shapes social reality. These self-fulfilling prophecies can have striking implications for the targets of those perceptions. The field of person perception has developed strong theoretical models of these processes, which imply change across time in both behavior and appearance. Thus, a combination of experimental and longitudinal observational methodologies are necessary to fully explore these models, which may have profound implications for other fields of psychology, such as development and psychopathology. In order to understand and interrupt negative self-fulfilling prophecies (e.g., those related to delinquency) and the development of an untrustworthy facial appearance, the experiences of both the perceiver and the target must be studied through methodological tools from both laboratory and longitudinal research.

Bibliography

Adams, G. R., & Cohen, A. S. (1974). Children's physical and interpersonal characteristics that effect student-teacher interactions. *The Journal of Experimental Education*, *43*(1), 1–5. https://doi.org/10.1080/00220973.1974.10806295

Alley, Z. M., Kerr, D. C. R., Wilson, J. P., & Rule, N. O. (2019). Prospective associations between boys' substance use and problem behavior histories and their facial trustworthiness in adulthood. *Journal of Social and Clinical Psychology*, *38*(8), 647–670. https://doi.org/10.1521/jscp.2019.38.7.647

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ : British Medical Journal*, *332*(7549), 1080.

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

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, *64*(3), 431–441. https://doi.org/10.1037/0022-3514.64.3.431

Babad, E. Y., Inbar, J., & Rosenthal, R. (1982). Teachers' judgment of students' potential as a function of teachers' susceptibility to biasing information. *Journal of Personality and Social Psychology*, *42*(3), 541–547. https://doi.org/10.1037/0022-3514.42.3.541

Bentler, P. M. (1990). *Comparative fit indexes in structural models.* Psychological Bulletin. https://doi.org/10.1037/0033-2909.107.2.238

Berry, D. S., & McArthur, L. Z. (1985). Some components and consequences of a babyface. *Journal of Personality and Social Psychology*, *48*(2), 312. https://doi.org/10.1037/0022-3514.48.2.312

Birkás, B., Dzhelyova, M., Lábadi, B., Bereczkei, T., & Perrett, D. I. (2014). Cross-cultural perception of trustworthiness: The effect of ethnicity features on evaluation of faces' observed trustworthiness across four samples. *Personality and Individual Differences*, *69*, 56–61. https://doi.org/10.1016/j.paid.2014.05.012

Capaldi, D. M., Chamberlain, P., Fetrow, R. A., & Wilson, J. E. (1997). Conducting ecologically valid prevention research: Recruiting and retaining a "Whole Village" in multimethod, multiagent studies. *American Journal of Community Psychology*, *25*(4), 471–492. https://doi.org/10.1023/A:1024607605690

Charlesworth, T. E. S., Hudson, S. T. J., Cogsdill, E. J., Spelke, E. S., & Banaji, M. R. (2019). Children use targets' facial appearance to guide and predict social behavior. *Developmental Psychology*. https://doi.org/10.1037/dev0000734

Chen, X. (2010). Desire for autonomy and adolescent delinquency: A latent growth curve analysis. *Criminal Justice and Behavior*, *37*(9), 989–1004. https://doi.org/10.1177/0093854810367481

Clifford, M. M., & Walster, E. (1973). The effect of physical attractiveness on teacher expectations. *Sociology of Education*, *46*(2), 248–258. JSTOR. https://doi.org/10.2307/2112099

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, *108*(1), 37–57. https://doi.org/10.1037/pspa0000014

Dare, G. J. (1992). The effect of pupil appearance on teacher expectations. *Early Child Development and Care*, *80*(1), 97–101. https://doi.org/10.1080/0300443920800112

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20–33. https://doi.org/10.1037/0022-3514.44.1.20

Dick, D. M., Adkins, A. E., & Kuo, S. I.-C. (2016). Genetic influences on adolescent behavior. *Neuroscience & Biobehavioral Reviews*, *70*, 198–205. https://doi.org/10.1016/j.neubiorev.2016.07.007

Dishion, T. J. (2000). Cross-setting consistency in early adolescent psychopathology: Deviant friendships and problem behavior sequalae. *Journal of Personality*, *68*(6), 1109–1126.

Dishion, T. J., McCord, J., & Poulin, F. (1999). When interventions harm: Peer groups and problem behavior. *The American Psychologist*, *54*(9), 755–764.

Dishion, T. J., & Owen, L. D. (2002). A longitudinal analysis of friendships and substance use: Bidirectional influence from adolescence to adulthood. *Developmental Psychology*, *38*(4), 480–491. https://doi.org/10.1037/0012-1649.38.4.480

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but…: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, *110*(1), 109–128. https://doi.org/10.1037/0033-2909.110.1.109

Elliott, D. S., Ageton, S. S., Huizinga, D., Knowles, B. A., & Canter, R. J. (1983). *The prevalence and incidence of delinquent behavior: 1976–1980—National Estimates of Delinquent Behavior by Sex, Race, Social Class and Other Selected Variables* (No. 26; National Youth Survey Report). Behavioral Research Institute. https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=128841

Flowe, H. D. (2012). Do characteristics of faces that convey trustworthiness and dominance underlie perceptions of criminality? *PLoS ONE*, *7*(6). https://doi.org/10.1371/journal.pone.0037253

Foo, Y. Z., Loncarevic, A., Simmons, L. W., Sutherland, C. A. M., & Rhodes, G. (2019). Sexual unfaithfulness can be judged with some accuracy from men's but not women's faces. *Royal Society Open Science*, *6*(4), 181552. https://doi.org/10.1098/rsos.181552

Friedrich, A., Flunger, B., Nagengast, B., Jonkmann, K., & Trautwein, U. (2015). Pygmalion effects in the classroom: Teacher expectancy effects on students' math achievement. *Contemporary Educational Psychology*, *41*, 1–12. https://doi.org/10.1016/j.cedpsych.2014.10.006

Giordano, P. C., Cernkovich, S. A., & Holland, D. D. (2003). Changes in friendship relations over the life course: Implications for desistance from crime. *Criminology*, *41*(2), 293–328. https://doi.org/10.1111/j.1745-9125.2003.tb00989.x

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, *83*(2), 314–320. https://doi.org/10.1037/0033-2909.83.2.314

Gunaydin, G., Selcuk, E., & Zayas, V. (2017). Impressions based on a portrait predict, 1-month later, impressions following a live interaction. *Social Psychological and Personality Science*, *8*(1), 36–44. https://doi.org/10.1177/1948550616662123

Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2019). Explaining the persistent influence of facial cues in social decision-making. *Journal of Experimental Psychology: General*, *148*(6), 1008–1021. https://doi.org/10.1037/xge0000591

Jessor, R. (1991). Risk behavior in adolescence: A psychosocial framework for understanding and action. *Journal of Adolescent Health*, *12*(8), 597–605. https://doi.org/10.1016/1054-139X(91)90007-K

King, R. D., & Light, M. T. (2019). Have racial and ethnic disparities in sentencing declined? *Crime and Justice*, *48*, 365–437. https://doi.org/10.1086/701505

Krumhuber, E., Manstead, A. S. R., Cosker, D., Marshall, D., Rosin, P. L., & Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion*, *7*(4), 730. https://doi.org/10.1037/1528-3542.7.4.730

Li, Q., Heyman, G. D., Mei, J., & Lee, K. (2017). Judging a book by its cover: Children's facial trustworthiness as judged by strangers predicts their real-world trustworthiness and peer relationships. *Child Development*, *90*(2). https://doi.org/10.1111/cdev.12907

Li, T., Liu, X., Pan, J., & Zhou, G. (2017). The interactive effect of facial appearance and behavior statement on trust belief and trust behavior. *Personality and Individual Differences*, *117*, 60–65. https://doi.org/10.1016/j.paid.2017.05.038

Little, A. C., Jones, B. C., DeBruine, L. M., & Dunbar, R. I. M. (2013). Accuracy in discrimination of self-reported cooperators using static facial information. *Personality and Individual Differences*, *54*(4), 507–512. https://doi.org/10.1016/j.paid.2012.10.018

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

Malatesta, C. Z., Fiore, M. J., & Messina, J. J. (1987). Affect, personality, and facial expressive characteristics of older people. *Psychology and Aging*, *2*(1), 64–69. https://doi.org/10.1037//0882-7974.2.1.64

Marcoulides, K. M. (2018). Automated latent growth curve model fitting: A segmentation and knot selection approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 687–699. https://doi.org/10.1080/10705511.2018.1424548

Marečková, K., Weinbrand, Z., Chakravarty, M. M., Lawrence, C., Aleong, R., Leonard, G., Perron, M., Pike, G. B., Richer, L., Veillette, S., Pausova, Z., & Paus, T. (2011). Testosterone-mediated sex differences in the face shape during adolescence: Subjective impressions and objective features. *Hormones and Behavior*, *60*(5), 681–690. https://doi.org/10.1016/j.yhbeh.2011.09.004

Muthen, B. O. (2008, August 5). Modeling with time-varying covariates. [Online forum comment]. Message posted to http://www.statmodel.com/discussion/messages/14/3460.html?1555003621

Okada, H. C., Alleyne, B., Varghai, K., Kinder, K., & Guyuron, B. (2013). Facial changes caused by smoking: A comparison between smoking and nonsmoking identical twins. *Plastic and Reconstructive Surgery*, *132*(5), 1085–1092. https://doi.org/10.1097/PRS.0b013e3182a4c20a

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

Ozono, H., Watabe, M., Yoshikawa, S., Nakashima, S., Rule, N. O., Ambady, N., & Reginald B. Adams, J. (2010). What's in a smile? Cultural differences in the effects of smiling on judgments of trustworthiness. *Letters on Evolutionary Behavioral Science*, *1*(1), 15–18. https://doi.org/10.5178/lebs.2010.4

Patterson, G. R. (1993). Orderly change in a stable world: The antisocial trait as a chimera. *Journal of Consulting and Clinical Psychology*, *61*(6), 911–919.

Patterson, G. R., DeBaryshe, B. D., & Ramsey, E. (1989). A developmental perspective on antisocial behavior. *The American Psychologist*, *44*(2), 329–335.

Porter, S., Brinke, L. ten, & Gustaw, C. (2010). Dangerous decisions: The impact of first impressions of trustworthiness on the evaluation of legal evidence and defendant culpability. *Psychology, Crime & Law*, *16*(6), 477–491. https://doi.org/10.1080/10683160902926141

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology*, *76*(1), 85–97. https://doi.org/10.1037/0022-0663.76.1.85

Rosenthal, R. (1987). Pygmalion effects: Existence, magnitude, and social importance. *Educational Researcher*, *16*(9), 37–40. https://doi.org/10.3102/0013189X016009037

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review*, *3*(1), 16–20. https://doi.org/10.1007/BF02322211

Rudasill, K. M., Reio, T. G., Stipanovic, N., & Taylor, J. E. (2010). A longitudinal study of student-teacher relationship quality, difficult temperament, and risky behavior from childhood to early adolescence. *Journal of School Psychology*, *48*(5), 389–412. https://doi.org/10.1016/j.jsp.2010.05.001

Runell, L. L. (2017). Identifying desistance pathways in a higher education program for formerly incarcerated individuals. *International Journal of Offender Therapy and Comparative Criminology*, *61*(8), 894–918. https://doi.org/10.1177/0306624X15608374

Sampson, R. J., & Laub, J. H. (2003). Life-course desisters? Trajectories of crime among delinquent boys followed to Age 70. *Criminology*, *41*(3), 555–592. https://doi.org/10.1111/j.1745-9125.2003.tb00997.x

Schaller, M. (2008). Evolutionary bases of first impressions. In N. Ambady & J. J. Skowronski (Eds.), *First impressions* (pp. 15–34). Guilford Publications.

Schulenberg, J. E., & Maggs, J. L. (2002). A developmental perspective on alcohol use and heavy drinking during adolescence and the transition to young adulthood. *Journal of Studies on Alcohol, Supplement*, *s14*, 54–70. https://doi.org/10.15288/jsas.2002.s14.54

Schweizer, K. (2010). Some guidelines concerning the modeling of traits and abilities in test construction. *European Journal of Psychological Assessment*, *26*(1), 1–2. https://doi.org/10.1027/1015-5759/a000001

Shen, X., Mann, T. C., & Ferguson, M. J. (2020). Beware a dishonest face?: Updating face-based implicit impressions using diagnostic behavioral information. *Journal of Experimental Social Psychology*, *86*, 103888. https://doi.org/10.1016/j.jesp.2019.103888

Skardhamar, T., & Savolainen, J. (2014). Changes in criminal offending around the time of job entry: A study of employment and desistance. *Criminology*, *52*(2), 263–291. https://doi.org/10.1111/1745-9125.12037

Snyder, M., Tanke, E. D., & Berscheid, E. (1977). Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes. *Journal of Personality and Social Psychology*, *35*(9), 656–666. https://doi.org/10.1037/0022-3514.35.9.656

Srull, T. K., & Wyer, R. S. (1981). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology*, *37*(10), 1660. https://doi.org/10.1037/0022-3514.37.10.1660

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences*, *108*(19), 7710–7715. https://doi.org/10.1073/pnas.1014345108

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502_4

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*(4), 298–303. https://doi.org/10.1016/j.evolhumbehav.2009.12.001

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: A model based approach. *Social Cognitive and Affective Neuroscience*, *3*(2), 119–127. https://doi.org/10.1093/scan/nsn009

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

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

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. https://doi.org/10.1007/BF02291170

Van Ryzin, M. J., & Dishion, T. J. (2014). Adolescent deviant peer clustering as an amplifying mechanism underlying the progression from early substance use to late adolescent dependence. *Journal of Child Psychology & Psychiatry*, *55*(10), 1153–1161. https://doi.org/10.1111/jcpp.12211

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

Wang, M.-T., Brinkworth, M., & Eccles, J. (2013). Moderating effects of teacher–student relationship in adolescent trajectories of emotional and behavioral adjustment. *Developmental Psychology*, *49*(4), 690–705. https://doi.org/10.1037/a0027916

Wiesner, M., & Capaldi, D. M. (2003). Relations of childhood and adolescent factors to offending trajectories of young men. *Journal of Research in Crime and Delinquency*, *40*(3), 231–262. https://doi.org/10.1177/0022427803253802

Wiesner, M., Kim, H. K., & Capaldi, D. M. (2005). Developmental trajectories of offending: Validation and prediction to young adult alcohol use, drug use, and depressive symptoms. *Development and Psychopathology*, *17*(1), 251–270.

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science*, *26*(8), 1325–1331. https://doi.org/10.1177/0956797615590992

Wilson, J. P., & Rule, N. O. (2016). Hypothetical sentencing decisions are associated with actual capital punishment outcomes: The role of facial trustworthiness. *Social Psychological and Personality Science*, *7*(4), 331–338. https://doi.org/10.1177/1948550615624142

Wilson, J. P., & Rule, N. O. (2017). Advances in understanding the detectability of trustworthiness from the face: Toward a taxonomy of a multifaceted construct. *Current Directions in Psychological Science*, *26*(4), 396–400. https://doi.org/10.1177/0963721416686211

Wu, Y., Gao, L., Wan, Y., Wang, F., Xu, S., Yang, Z., Rao, H., & Pan, Y. (2018). Effects of facial trustworthiness and gender on decision making in the Ultimatum Game. *Social Behavior and Personality: An International Journal*, *46*(3), 499–516. https://doi.org/doi.org/10.2224/sbp.6966

Zebrowitz, L. A. (1997). The Bases of Reading Faces by Zebrowitz. In *Reading faces: Window to the soul?* (pp. 40–63). Routledge.

Zebrowitz, L. A., Andreoletti, C., Collins, M. A., Lee, S. Y., & Blumenthal, J. (1998). Bright, bad, babyfaced boys: Appearance stereotypes do not always yield self-fulfilling prophecy effects. *Journal of Personality and Social Psychology*, *75*(5), 1300–1320. https://doi.org/10.1037//0022-3514.75.5.1300

Zebrowitz, L. A., & McDonald, S. M. (1991). The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior*, *15*(6), 603–623. https://doi.org/10.1007/BF01065855

Zebrowitz, L. A., & Montepare, J. (2006). The ecological approach to person perception: Evolutionary roots and contemporary offshoots. In *Evolution and social psychology* (pp. 81–113). Psychosocial Press.

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass*, *2*(3), 1497. https://doi.org/10.1111/j.1751-9004.2008.00109.x

Zebrowitz, L. A., Olson, K., & Hoffman, K. (1993). Stability of babyfaceness and attractiveness across the life span. *Journal of Personality and Social Psychology*, *64*(3), 453–466. https://doi.org/10.1037/0022-3514.64.3.453

APPENDICES

Appendix A: Dependent Measures

Personality Assessment

Think about your impression of this man's personality, based on the information you read in the previous vignette. Please report your impression of him on the following traits.

There are no right answers; just give your best impression of that person based on the information you have.

How *HOSTILE* is he?

1	2	3	4	5	6	7
Not at all hostile						Extremely hostile

How *TRUSTWORTHY* is he?

1	2	3	4	5	6	7
Not at all trustworthy						Extremely trustworthy

How *INTELLIGENT* is he?

1	2	3	4	5	6	7
Not at all intelligent						Extremely intelligent

How *FRIENDLY* is he?

1	2	3	4	5	6	7
Not at all friendly						Extremely friendly

How *LIKEABLE* is he?

1	2	3	4	5	6	7
Not at all likeable						Extremely likeable

How *CONSIDERATE* is he?

1	2	3	4	5	6	7
Not at all considerate						Extremely considerate

How *KIND* is he?

1	2	3	4	5	6	7
Not at all kind						Extremely kind

How *AGGRESSIVE* is he?

1	2	3	4	5	6	7
Not at all aggressive						Extremely aggressive

How *THOUGHTFUL* is he?

1	2	3	4	5	6	7
Not at all thoughtful						Extremely thoughtful

Behavioral Intent Measure

Think about your impression of this man, based on the information you read in the previous vignette. Please report how you would be likely to respond to him yourself in the contexts described below. There are no right answers, so please give your best guess based on the information that you do have.

ITEM 1: If you had a choice, how likely would you be to work with him on a group project?

1	2	3	4	5	6	7
Not at all likely						Extremely likely

ITEM 2: If you were hiring for a job, and this person was a qualified candidate, how likely would you be to hire him?

1	2	3	4	5	6	7
Not at all likely						Extremely likely

ITEM 3: If you had a choice, how likely would you be to choose this person as a partner on a project for work?

1	2	3	4	5	6	7
Not at all likely						Extremely likely

ITEM 4: If you saw him at a party, how likely would you be to talk to him?

1	2	3	4	5	6	7
Not at all likely						Extremely likely

ITEM 5: If you were sitting in a waiting room with him, how likely would you be to chat with him?

1	2	3	4	5	6	7
Not at all likely						Extremely likely

ITEM 6: If you met this person casually and he sent you a friend request on social media, how likely would you be to accept?

1	2	3	4	5	6	7
Not at all likely						Extremely likely

ITEM 7 (Attention Check): If you are reading this survey, would you please select 7 – Extremely likely for this item only, and continue to answer the other items normally?

1	2	3	4	5	6	7
Not at all likely						Extremely likely

ITEM 8: If you wanted to step away from your backpack at an airport, how willing would you be to ask him to watch it for you while you're gone?

1	2	3	4	5	6	7
Not at all likely						Extremely likely

ITEM 9: If this person was your friend and asked you to lend him $50, how likely would you be to lend him the money?

1	2	3	4	5	6	7
Not at all likely						Extremely likely

ITEM 10: If you saw this person standing around in your neighborhood for long periods of time, how likely would you be to report him as a suspicious person?

1	2	3	4	5	6	7
Not at all likely						Extremely likely

ITEM 11: If you were a police officer and you pulled this man over for driving 65mph in a 55mph zone, how likely would you be to give him a ticket (versus a warning)?

1	2	3	4	5	6	7
Not at all likely						Extremely likely

ITEM 12: If you heard that this person was under investigation for petty theft (an item less than $100), how confident would you be that this person had committed the crime?

1	2	3	4	5	6	7
Not confident at all						Very confident

ITEM 13: If you were a judge and this person was found guilty of petty theft (an item less than $100, and with no prior convictions), how much do you think this person should be fined?

1	2	3	4	5	6	7
$400			$1200			$1800

ITEM 14: If you heard that this person was under investigation for stealing a car, how confident would you be that this person had committed the crime?

1	2	3	4	5	6	7
Not confident at all						Very confident

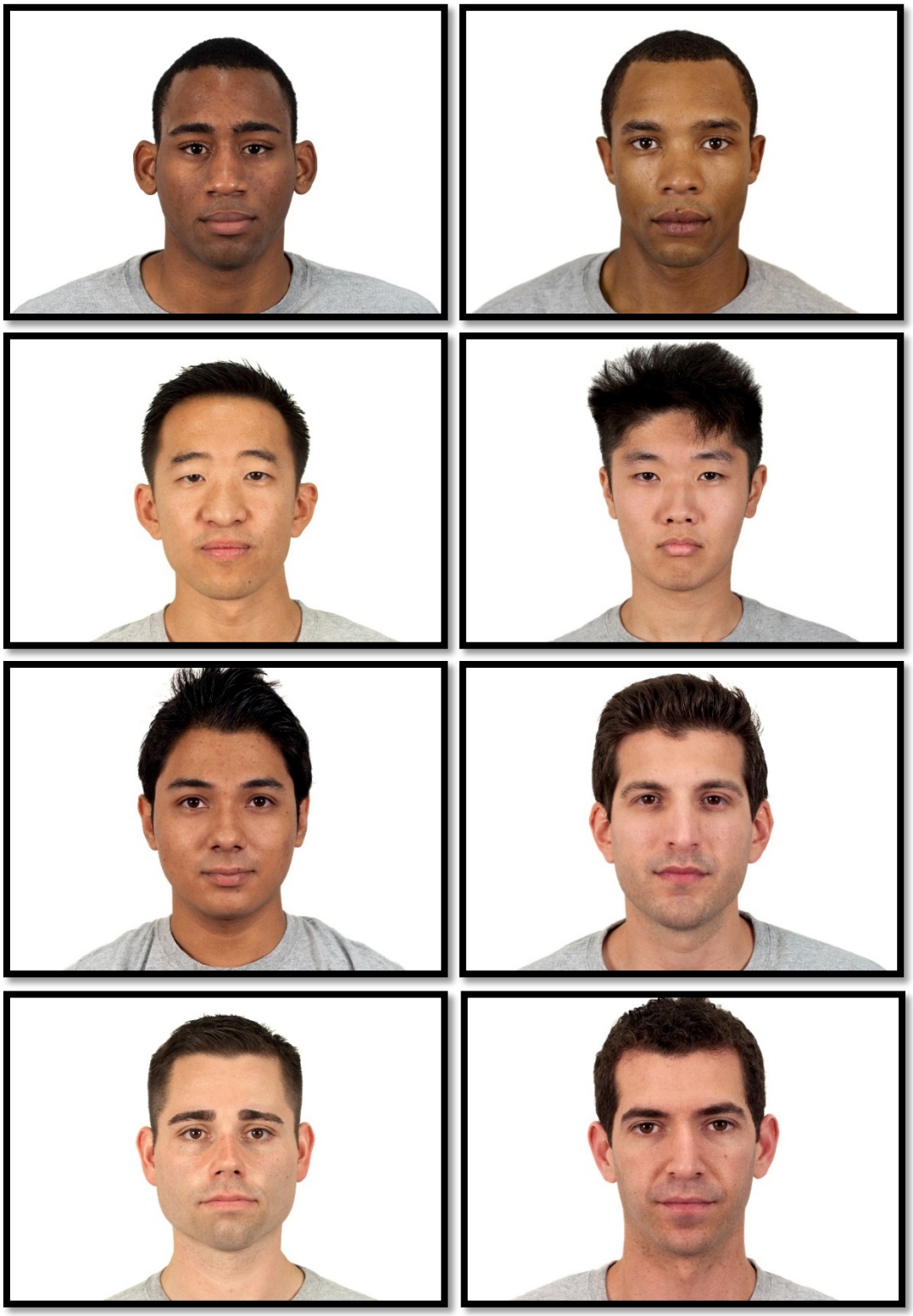
ITEM 15: If you were a judge and this person was found guilty of stealing a car (with no prior convictions), how much time would you sentence him to serve in jail or prison?

1	2	3	4	5	6	7
16 months			2 years			3 years

Appendix B: Facial Stimuli

Untrustworthy Faces

Trustworthy Faces



Appendix C: Behavioral Vignettes

Vignettes marked with an asterisk were selected for Study 2.

Original "Donald" Vignette (Srull & Wyer, 1981)

I ran into my old acquaintance James the other day, and I decided to go over and visit him, since by coincidence we took our vacations at the same time. Soon after I arrived, a salesman knocked at the door, but James refused to let him enter. He also told me that he was refusing to pay his rent until the landlord repaints his apartment. We talked for a while, had lunch, and then went out for a ride. We used my car, since James's car had broken down that morning, and he told the garage mechanic that he would have to go somewhere else if he couldn't fix his car that same day. We went to the park for about an hour and then stopped at a hardware store. I was sort of preoccupied, but James bought some small gadget, and then I heard him demand his money back from the sales clerk. I couldn't find what I was looking for, so we left and walked a few blocks to another store. The Red Cross had set up a stand by the door and asked us to donate blood. James lied by saying he had diabetes and therefore could not give blood. It's funny that I hadn't noticed it before, but when we got to the store, we found that it had gone out of business. It was getting kind of late, so I took James to pick up his car and we agreed to meet again as soon as possible.

**Assertive Vignette (based on the "Donald" vignette)*

I ran into my old acquaintance Joseph recently, and we both agreed to hang out the next day. Joseph's car broke down on his way to see me, so we towed it to a nearby mechanic. The mechanic said it would take a week to repair, but Joseph said he would take his car elsewhere if the garage couldn't fix it the same day. We had lunch together

and talked for a while. I mentioned that my car had been acting funny too, so we decided to walk to a nearby auto shop to get a part. When we got to the shop, I couldn't find what I was looking for, but Joseph bought something small. He noticed it was missing a component just before we left, and he demanded his money back even though the sign said "No refunds." On our way to a local pub, we passed by a charity stand asking for donations. I didn't have any cash, and Joseph said he wasn't interested when they flagged us down. A salesperson approached us before we got to the pub. Joseph refused to listen to his pitch, but the salesperson didn't leave until I said we didn't have any cash. We got to the pub and talked for a while. Joseph had an issue with the flooring in his apartment, and he was refusing to pay his rent until his landlord agreed to replace it. I gave Joseph a ride back to the garage, and we decided to get together again next week.

Passive Vignette (based on the "Donald" vignette)

I ran into my old acquaintance John the other day, and I decided to go over and visit him, since we happened to have taken our vacations at the same time. Soon after I arrived, a salesman knocked at the door. John listened to his pitch for about twenty minutes. Afterward, John told me he was hoping that his landlord would notice the chipping paint on his walls. We had lunch and then went out for a ride. We used my car, since John's car would be getting repairs for the next week. John said he wished he had taken it to a different shop, since he was sure the repairs could have been done in a day. We went to the park for about an hour and then stopped at a hardware store. I was sort of preoccupied, but John bought some small gadget. He went up to the salesclerk because there was an issue with the gadget, but he didn't protest when the clerk told him "No refunds." I couldn't find what I was looking for, so we left and walked a few blocks to

another store. The Red Cross had set up a stand by the door and asked us to donate blood. John seemed uncomfortable, but he agreed to donate blood. It's funny that I hadn't noticed it before, but when we got to the store, we found that it had gone out of business. It was getting kind of late, so I dropped John off at his apartment and we agreed to meet again as soon as possible.

Passive Vignette (based on the "Assertive Vignette")

I ran into my old acquaintance Jack recently, and we both agreed to hang out the next day. Jack's car broke down on his way to see me, so we towed it to a nearby mechanic. The mechanic said it would take a week to repair, but after we left Jack told me a different garage could have done it in a day. We had lunch together and talked for a while. I mentioned that my car had been acting funny too, so we decided to walk to a nearby auto shop to get a part. When we got to the shop, I couldn't find what I was looking for, but Jack bought something small. He noticed it was missing a component just before we left, but he didn't protest when the salesclerk told him "No refunds." On our way to a local pub, we passed by a charity stand asking for donations. I didn't have any cash, but Jack donated some money when they flagged us down. A salesperson approached us before we got to the pub. Jack listened to his pitch for about 20 minutes, but the salesperson left when I said we didn't have any cash. We got to the pub and talked for a while. Jack had an issue with the flooring in his apartment, but he was waiting for the right time to talk to his landlord about it. I gave Jack a ride back to his apartment, and we decided to get together again next week.

Appendix D: Behavioral Intent Factor Analysis

Rotated Component Matrix for Assertive Condition

	Component			
	1	2	3	4
ITEM 1	0.890	0.173	0.083	-0.050
ITEM 2	0.818	0.249	0.043	0.127
ITEM 3	0.893	0.219	0.052	-0.078
ITEM 4	0.412	0.791	-0.034	0.008
ITEM 5	0.330	0.853	0.023	0.085
ITEM 6	0.183	0.780	0.099	-0.184
ITEM 8	0.556	0.420	0.192	0.073
ITEM 9	0.591	0.361	0.344	-0.117
ITEM 10	0.250	-0.275	0.611	0.270
ITEM 11	0.011	0.085	0.358	0.620
ITEM 12	0.047	0.093	0.891	0.101
ITEM 13	-0.217	0.022	0.056	0.788
ITEM 14	0.090	0.157	0.856	0.101
ITEM 15	0.185	-0.178	0.054	0.785

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
See Appendix A for items. Note, ITEM 7 is an attention check and therefore was not included in the factor analysis.

Rotated Component Matrix for Passive Condition

	Component			
	1	2	3	4
ITEM 1	0.266	0.869	-0.024	0.041
ITEM 2	0.202	0.865	-0.126	-0.002
ITEM 3	0.317	0.851	0.111	0.022
ITEM 4	0.814	0.292	0.037	0.042
ITEM 5	0.832	0.063	0.025	-0.018
ITEM 6	0.806	0.091	-0.072	0.090
ITEM 8	0.702	0.222	0.054	0.050
ITEM 9	0.507	0.224	0.028	-0.005
ITEM 10	-0.004	0.097	0.053	0.810
ITEM 11	0.056	-0.039	0.176	0.764
ITEM 12	0.249	-0.077	0.676	0.292
ITEM 13	-0.096	-0.064	0.781	-0.176
ITEM 14	0.217	-0.130	0.685	0.325
ITEM 15	-0.222	0.253	0.604	0.121

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
See Appendix A for items. Note, ITEM 7 is an attention check and therefore was not included in the factor analysis.

Final Factor Structure

	Component			
	Responsibility	Social	Suspicion	Punishment
ITEM 1	X			
ITEM 2	X			
ITEM 3	X			
ITEM 4		X		
ITEM 5		X		
ITEM 6		X		
ITEM 8	X			
ITEM 9	X			
ITEM 10			X	
ITEM 11				X
ITEM 12			X	
ITEM 13				X
ITEM 14			X	
ITEM 15				X

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
See Appendix A for items. Note, ITEM 7 is an attention check and therefore was not included in the factor analysis.