

## Tilburg University

### Facial discrimination

Jaeger, B.

*Publication date:*  
2020

*Document Version*  
Publisher's PDF, also known as Version of record

[Link to publication in Tilburg University Research Portal](#)

*Citation for published version (APA):*  
Jaeger, B. (2020). *Facial discrimination: The irresistible influence of first impressions*. Proefschriftmaken.

#### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

#### Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# FACIAL DISCRIMINATION

The irresistible influence  
of first impressions

---

Bastian Jaeger



# **Facial Discrimination**

The irresistible influence of first impressions

**Bastian Jaeger**

Design by: Floor Weijs || [www.floorweijs.nl](http://www.floorweijs.nl)

Printed by: ProefschriftMaken || [www.proefschriftmaken.nl](http://www.proefschriftmaken.nl)

© 2020 Bastian Jaeger

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage or retrieval system, without permission in writing from the author. The copyright of the articles that have been accepted for publication or that already have been published, has been transferred to the respective journals.

# Facial Discrimination

The irresistible influence of first impressions

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan Tilburg University

op gezag van de rector magnificus,

prof. dr. K. Sijtsma,

in het openbaar te verdedigen

ten overstaan van een door het college

voor promoties aangewezen commissie

in de Aula van de Universiteit

op vrijdag 27 maart 2020 om 10.00

door Bastian Jaeger,

geboren te Duisburg, Duitsland.

**Promotor**

Prof. Dr. Ilja van Beest

**Copromotores**

Dr. Anthony M. Evans

Dr. Mariëlle Stel

**Promotiecommissie**

Prof. Dr. Alexander T. Todorov

Prof. Dr. Detlef Fetchenhauer

Prof. Dr. Ellen Giebels

Dr. Claudia Toma

Dr. Mariska E. Kret

# Contents

Chapter 1	Introduction	5
	First impressions from facial features	10
	First impressions as social biases	18
	Outline of the dissertation	19

## **PART I: The influence of first impressions**

Chapter 2	Facial appearance and electoral success: Are trustworthy-looking politicians more successful in corrupt regions?	23
	Study 2.1 - 2.3	28
Chapter 3	The effects of facial attractiveness and trustworthiness in online peer-to-peer markets	51
	Study 3.1	56

## **PART II: Why people rely on first impressions**

Chapter 4	Who judges a book by its cover? The prevalence, structure, and correlates of beliefs in physiognomy	77
	Study 4.1 & 4.2	84
	Study 4.3	90
	Study 4.4	98
	Study 4.5	104

Chapter 5	The bounds of physiognomy: Lay beliefs in the manifestation of personality traits in facial features	117
	Study 5.1	124
	Study 5.2	130
	Study 5.3	135
Chapter 6	Explaining the persistent influence of facial cues in social decision-making	147
	Study 6.1	154
	Study 6.2	156
	Study 6.3 – 6.5	164
	Study 6.6	170
<b>Part III: Mitigating the influence of first impressions</b>		
Chapter 7	Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions	183
	Study 7.1	190
	Study 7.2	196
	Study 7.3	202
Chapter 8	General discussion	213
	Summary of main findings	214
	Theoretical implication	218
	Practical implications	226
	Limitations and open questions	230
	Future directions for the study of first impressions	235
	Conclusion	239



Appendix	Automated classification of demographics from face images: A tutorial and validation	241
	Study A.1	250
	Study A.2	261
	References	279
	Acknowledgments	321
	Kurt Lewin Institute dissertation series	325



# Chapter 1

Introduction

## Chapter 1

On 27 December 1831, Charles Darwin left Plymouth harbor on board of the HMS Beagle. As the naturalist of the expedition, he made countless observations, which would later culminate in his theory of evolution—arguably, one of the most influential scientific theories of all time. However, as Darwin recounts in his autobiography, he was almost prevented from boarding the Beagle due to the shape of his nose (Darwin, 1887). The ship’s captain was a firm believer in physiognomy, which posits that a person’s facial appearance contains information about their character (Lavater, 1775). He was convinced that nobody with such a nose could have sufficient stamina to take part in the long and arduous journey. Luckily, Darwin was allowed to board the Beagle in the end. Later, he somewhat dryly remarked that the captain “was afterwards well satisfied that my nose had spoken falsely”.

To the scientifically-inclined reader, the captain’s behavior may seem comically irrational. How could the shape of a nose reveal something about a person’s character? However, the captain was not alone in holding this belief. The central idea of physiognomy—that stable, morphological features of a faces are indicative of character traits—has a long history in scholarly thought. Physiognomic writings date back to at least the time of Ancient Greece (Aristotle, trans. 1936) and the idea was particularly prominent in the 18<sup>th</sup> and 19<sup>th</sup> century (Lavater, 1775; Woods, 2017). However, physiognomic claims were often vague, inconsistent, and not based on scientific study (Alley, 1988; Collins, 1999). In fact, empirical tests of proposed relationships between specific morphological features of faces and psychological characteristics at the beginning of the 20<sup>th</sup> century yielded no support (Cleeton & Knight, 1924). Today, physiognomy is widely regarded as pseudo-science (Todorov, 2017).

Yet, research in the field of social perception has shown that physiognomic judgments are pervasive in everyday life: People spontaneously judge another person’s character based on their facial

features (Freeman & Johnson, 2016; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015).<sup>1</sup> Stereotypes regarding what a trustworthy or dominant person looks like are widely shared (Hehman, Sutherland, Flake, & Slepian, 2017; Oosterhof & Todorov, 2008; Rule, Krendl, Ivcevic, & Ambady, 2013) and they are triggered within a few hundred milliseconds of perceiving a face (Todorov, Pakrashi, & Oosterhof, 2009; Willis & Todorov, 2006). Moreover, people are relatively confident in the accuracy of their trait impressions (Ames, Kammrath, Suppes, & Bolger, 2010) and rely on them when making a wide range of decisions (Olivola, Funk, & Todorov, 2014).

The widespread influence of split-second personality judgments based on facial features may seem surprising given people's limited ability to infer personality from facial appearance (Bonnenfon, Hopfensitz, & De Neys, 2017; Todorov, Olivola, et al., 2015). The question of whether there is any accuracy in trait impressions (and what mechanisms might account for this) remains heavily debated (Bonnenfon, Hopfensitz, & De Neys, 2015; Efferson & Vogt, 2013; Todorov, Funk, & Olivola, 2015). However, there is ample evidence that people *overrely* on facial appearance. That is, facial cues exert a disproportionate influence on decisions given their low diagnosticity for inferring personality traits and behavioral tendencies. For example, people rely on facial appearance even when making consequential decisions (e.g., criminal sentencing decisions; J. P. Wilson & Rule, 2015), even when they have better information is available (Olivola, Tingley, & Todorov, 2018), and even when they are explicitly told to discount a person's appearance (Blair, Judd, & Fallman, 2004; Hassin & Trope, 2000). Moreover, people are often biased in favor of individuals with an attractive facial appearance: Impressions of attractiveness influence many judgments

---

<sup>1</sup> In this dissertation, I focus on facial appearance, rather than other stimuli such as clothing or hairstyle, because faces are arguably the most relevant stimulus in human signaling and impression formation (Hugenberg & Wilson, 2013; Leopold & Rhodes, 2010).

## Chapter 1

and behaviors, even in situations where attractiveness should not play a role (Maestriperieri, Henry, & Nickels, 2017).

While the effects of facial appearance are well-documented, little is known about the underlying mechanisms. Why do people persistently rely on first impressions? I argue that in order to answer this question, we should treat and study first impressions more like other social biases.<sup>2</sup> Similar to discrimination based on gender, race, or sexual orientation, discrimination based on facial appearance unfairly disfavors a certain individuals. For example, as a consequence of their facial appearance, individuals might be less likely to be promoted (Ling, Luo, & She, 2019), granted a loan (Duarte, Siegel, & Young, 2012), or trusted by others (van't Wout & Sanfey, 2008); they might receive a lower wage at work (Hamermesh & Biddle, 1994), less help from others (Maestriperieri et al., 2017), and harsher sentences in court (J. P. Wilson & Rule, 2015; Zebrowitz & McDonald, 1991). In short, facial discrimination is pervasive. Some evidence even suggests that correcting facial stereotypes may be more difficult than correcting gender or race stereotypes (Blair et al., 2004; Walker et al., 2017).

Treating first impressions as a biasing influence on social behavior and decision-making is not novel (Maestriperieri et al., 2017; Olivola, Funk, et al., 2014; Zebrowitz, Fellous, Mignault, & Andreoletti, 2003). However, a social bias perspective focuses the study of first impressions on theoretically and practically important questions that have not received much attention thus far. Research on social biases often follows three general goals: (a) demonstrating the existence of the bias and its pervasiveness (e.g., in which domains does racial discrimination occur?), (b) understanding its underlying mechanisms (e.g., why do people discriminate based on race?), and (c) developing interventions to curb

---

<sup>2</sup> Social bias is generally defined as “intended or unintended favoritism in evaluation, judgment, or behavior for one social group over another” (Axt & Nosek, 2018, p. 337).

the bias (e.g., how can we reduce racial biases?). Crucially, prior research on first impressions has predominantly focused on the first issue, documenting the various ways in which facial appearance influences social outcomes. Little is known about why people rely on first impressions or how this can be prevented. In this dissertation, I aim to address these gaps in the literature.

A social bias perspective also suggests that hypotheses about the underlying causes of facial discrimination can be derived from existing theories on social biases. For example, bias taxonomies in judgment and decision-making offer insights into common sources of biased behavior (T. D. Wilson & Brekke, 1994) and may help elucidate why people persistently rely on first impressions. In a similar vein, years of social-psychological research on explicit and implicit bias has resulted in a long list of interventions that may not only change racial biases (Forscher et al., 2019; Lai et al., 2014), but also reliance on facial stereotypes. In short, this dissertation attempts a closer integration of the first impression literature with other research areas that are concerned with understanding and alleviating social biases.

Ultimately, I hope that this approach will advance our understanding of when people discriminate based on facial features, what the underlying cognitive mechanisms are, and how we can reduce facial discrimination. Before describing the work I conducted to tackle these questions, I will provide a brief review of the first impression literature. Understanding how people process faces is crucial for understanding the downstream consequences of first impressions. I will focus on how people form trait impressions based on facial features, how accurate their impressions are, and how their impressions influence decision-making. I will then outline how a social bias perspective raises a number of central, but largely unaddressed research questions regarding the influence of first impressions. Finally, I will describe how

each of these questions is tackled in the six empirical papers that are reported in this dissertation.

### **First impressions from facial features**

The human face is a rich source of information. Morphological features provide information about a person's identity, sex, age, and race (Bruce & Young, 2012) and dynamic features, such as emotion expressions, can signal current intentions or feelings (Crivelli & Fridlund, 2018; Van Kleef, 2010). The relevance of facially communicated information for social interaction is underscored by the existence of specialized perceptual and cognitive systems for processing faces (Farah, Wilson, Drain, & Tanaka, 1998; Hugenberg & Wilson, 2013; Palermo & Rhodes, 2007). Faces attract attention (Ro, Russell, & Lavie, 2001; Theeuwes & Van der Stigchel, 2006) and this is particularly true for faces conveying socially relevant information (e.g., expressive faces; Pinkham, Griffin, Baron, Sasson, & Gur, 2010). Even newborns preferentially orient themselves towards faces, suggesting that the cognitive architecture for detecting and processing faces is already present at birth (Farroni et al., 2005). Moreover, humans can detect even slight changes in facial skin coloration (Thorstenson & Elliot, 2017) or facial expression (Leleu et al., 2018). In sum, interpreting faces is crucial for navigating the social world. For this reason, information communicated by faces is processed in a very quick and efficient manner.

### **Forming trait impressions from faces**

People not only infer demographic features or affective states from faces, but also a variety of traits and dispositions (Todorov, Olivola, et al., 2015). That is, depending on the morphological structure of their face, a person may be perceived as trustworthy, outgoing, or intelligent. Research on the content and structure of trait impressions has revealed that faces are primarily judged along three core dimensions (Oosterhof



& Todorov, 2008; Sutherland et al., 2013; Sutherland, Rhodes, Burton, & Young, 2019; see also B. C. Jones et al., 2019): intentions (e.g., trustworthiness, approachability, sociability), abilities (e.g., dominance, competence, intelligence), and youthfulness-attractiveness (e.g., attractiveness, age, health). Judgments of intentions and abilities are thought to reflect evaluations of a target's threat potential (Oosterhof & Todorov, 2008). In a similar vein, attractiveness judgments are thought to reflect the evaluation of a person's mate value (Rhodes, 2006; Rhodes, Simmons, & Peters, 2005). Thus, faces are judged along three, largely independent dimensions. These dimensions are best approximated by impressions of trustworthiness, dominance, and attractiveness and reflect evaluations of a person's perceived value as a social or sexual partner.

Trait impressions are formed on the basis of various facial characteristics. People rely on morphological features, such as facial width-to-height ratio (Hehman, Leitner, & Gaertner, 2013; Stirrat & Perrett, 2010), symmetry, sexual dimorphism, and averageness (A. L. Jones & Jaeger, 2019; Oosterhof & Todorov, 2008; Rhodes et al., 2007), but also on skin texture (Jaeger, Wagemans, Evans, & van Beest, 2018), color (Thorstenson, Pazda, Elliot, & Perrett, 2017), and contrast (Russell et al., 2016). A wealth of studies has identified There are many systematic relationships between specific facial characteristics and trait impressions (e.g., feminine features are seen as trustworthy), since people rely on similar characteristics when judging specific traits. Even though individual differences undoubtedly exist (Hehman et al., 2017; Hönekopp, 2006), significant consensus in trait impressions is found across different cultures (Rule et al., 2010) and age groups (Cogsdill & Banaji, 2015; Zebrowitz & Montepare, 1992).

Converging evidence suggests that the evaluation of faces on social dimensions is a largely automatic process. Trait judgments are formed within a few hundred milliseconds of exposure to a face (Todorov et al.,

2009; Willis & Todorov, 2006). These judgments are formed spontaneously, that is, even when engaging in a task that does not require character evaluations (Engell, Haxby, & Todorov, 2007; Klapper, Dotsch, van Rooij, & Wigboldus, 2016). Some evidence even suggests that trait impressions are formed before humans are consciously aware of perceiving a face (Hung, Nieh, & Hsieh, 2016; Stewart et al., 2012; Winston, Strange, O'Doherty, & Dolan, 2002). As a consequence, forming first impressions is often described as a mandatory (Ritchie, Palermo, & Rhodes, 2017), reflexive (Tabak & Zayas, 2012), and instantaneous process (Freeman & Johnson, 2016).

### **Accuracy of trait impressions**

How accurate are split-second judgments based on facial features? Some have argued that personality judgments on the basis of facial appearance (i.e., based on static images of resting, non-expressive faces) contain a “kernel of truth” (Berry, 1990; Bonnefon, Hopfensitz, & De Neys, 2013; Penton-Voak, Pound, Little, & Perrett, 2006). This view holds that some facial cues co-occur with personality traits and that people rely on them to form accurate impressions. However, a critical examination of the literature reveals that evidence for this view is not particularly strong.

First, empirical evidence in support of the kernel of truth hypothesis is inconsistent. One stream of research has focused on whether people can predict behavioral tendencies (e.g., trustworthiness, corruptibility) based on facial appearance. For example, in one commonly adopted design, participants play an economic game in which they decide whether to trust an interaction partner based on a facial photograph (Bonnefon et al., 2013). Across multiple studies, Bonnefon and colleagues (Bonnefon et al., 2013, 2017; De Neys, Hopfensitz, & Bonnefon, 2017) found that participants were more likely to engage with partners who tended to reciprocate, rather than betray their trust at

rates slightly above chance (ca. 55%). These results suggest that participants were able to identify trustworthy interaction partners based on their facial features. However, a review of the literature shows that evidence in favor of this conclusions is mixed and many studies do not find any accuracy in trustworthiness detection (Efferson & Vogt, 2013; C. Lin, Adolphs, & Alvarez, 2018; Rule et al., 2013; Slepian & Ames, 2015; Sylwester, Lyons, Buchanan, Nettle, & Roberts, 2012; Tognetti, Berticat, Raymond, & Faurie, 2013; Vogt, Efferson, & Fehr, 2013). Moreover, different investigations suggest that accuracy depends on a variety of different factors, such as the target's sex (Tognetti et al., 2013), at what moment the photo was taken (Verplaetse, Vanneste, & Braeckman, 2007), or how personality evaluations are measured (behavioral trust vs. trustworthiness ratings; Bonnefon et al., 2013). Together, these results cast doubt on the conclusion that people can reliably detect trustworthiness from facial features.

Two related approaches have yielded similarly inconsistent results. For example, researchers have examined correlations between self-reported Big Five personality traits and perceived personality traits based on facial photographs (Borkenau, Brecke, Möttig, & Paelecke, 2009). Another common approach is to select individuals that score particularly high or low on a certain dimension (e.g. extraversion) and morph their faces in order to create extraverted and introverted face prototypes (Penton-Voak et al., 2006). Participants then choose which of the two face composites scores higher on the dimension of interest.

For judgments of extraversion, which usually show the highest levels of accuracy in stranger rating tasks (Kenny & West, 2008), evidence for the kernel of truth hypothesis is mixed. While some studies find significant levels of accuracy (Borkenau et al., 2009; Kramer & Ward, 2010; Naumann, Vazire, Rentfrow, & Gosling, 2009; Penton-Voak et al., 2006), others do not (Ames et al., 2010; A. L. Jones, Kramer, & Ward, 2012; Shevlin, Walker, Davies, Banyard, & Lewis, 2003). Similarly

## Chapter 1

patterns were found for judgments of openness, conscientiousness, agreeableness, and neuroticism. Overall, evidence for accurate detection of Big Five personality traits from facial features is inconsistent—even when differences in facial features are exaggerated by morphing images of people who score highest and lowest on a certain domain (Little & Perrett, 2007; Penton-Voak et al., 2006).

There are a number of additional reasons that speak against accuracy in trait impressions. If judgments reflect the detected personality of a specific individual that is shown in a photo, then judgments of the same individual should be relatively consistent. However, trait judgments vary substantially across different photos of the same individual (Todorov & Porter, 2014a), different perceivers (Hehman et al., 2017), and different contexts (Brambilla, Biella, & Freeman, 2018), suggesting that they do not reflect reliable evaluations of an individual's personality.

In addition to weak support for accuracy in personality judgments, it is unclear which mechanisms could account for a relationship between facial appearance and personality traits. Some have proposed that discrimination based on facial features might result in a self-fulfilling prophecy (Slepian & Ames, 2015). For example, people who are (at first incorrectly) perceived as trustworthy due to their facial appearance might be treated in such a way that they actually turn out to be untrustworthy. This view predicts that people with similar faces should develop similar personality traits because they are treated in similar ways. Yet, studies with genetically unrelated individuals with high levels of facial similarity have yielded no support for this prediction (Segal, 2013; Segal, Graham, & Ettinger, 2013; Segal, Hernandez, Graham, & Ettinger, 2018).

A different theory holds that facial width-to-height ratio and personality are influenced by common biological factors (Deaner, Goetz, Shattuck, & Schnotala, 2012; Stirrat & Perrett, 2010). However, a recent

study found little support for relationships between facial width-to-height ratio and a wide range of psychological attributes (Kosinski, 2017). In sum, theoretical and empirical support for the idea that personality traits can be reliably inferred from facial appearance is weak and inconsistent. While more rigorous, high-powered studies are needed to address which traits may be reflected in facial features (and what can account for this), the available evidence suggests that people's ability to infer personality traits from faces is weak at best.

### **Functional significance of trait impressions**

What are the ultimate explanations for people's tendency to form trait impressions from facial features? The current evidence suggests that trait impressions are generally inaccurate, so why do people form them in the first place? From an evolutionary point of view, automatic but inaccurate inference may seem paradoxical.

The most convincing explanation is provided by overgeneralization theory (Zebrowitz, 2017; Zebrowitz et al., 2003). Even slight resemblances between facial features and socially relevant stimuli are sufficient to trigger associated stereotypes and responses. For example, facial features that slightly resemble emotion expressions (e.g., lowered eyebrows resembling an angry scowl or upturned corners of the mouth resembling a smile) can elicit inferences that are congruent with these emotional states (e.g., a smiling person is trustworthy and friendly; Adams, Nelson, Soto, Hess, & Kleck, 2012; Said, Sebe, & Todorov, 2009). In a similar vein, feminine facial features trigger trait impressions congruent with gender stereotypes (Oh, Buck, & Todorov, 2019; Walker et al., 2017) and baby-ish features (i.e., large eyes and round faces) trigger trait impressions associated with infants (Zebrowitz & Montepare, 1992). Thus, first impressions are by-products of otherwise adaptive mechanisms that extract socially relevant information from facial appearance. This explains why people automatically and

## Chapter 1

consensually infer a variety of characteristics from facial appearance even though these judgments have very low or no predictive validity.

A functional view of first impressions is also supported by studies on non-human primates and other animals. If the cognitive mechanisms that give rise to first impressions have evolved due to selection pressures in humans' evolutionary past, then other species that faced similar pressures (or that have evolved from a common ancestor) should show similar reactions to facial cues. To test this hypothesis, researchers have examined how primates' gaze behavior is influenced by variations in facial features (rather than asking them to rate faces on a nine-point scale; Leopold & Rhodes, 2010). Rhesus macaques spent more time looking at the faces of conspecifics when their facial symmetry (Waitt & Little, 2006) or redness (Waitt et al., 2003) were increased. These results mirror findings from experiments in which faces with increased symmetry (Perrett et al., 1999; Rhodes et al., 2007) and redness (Han et al., 2017; Thorstenson & Elliot, 2017; Thorstenson et al., 2017) were rated as more attractive by human subjects. These findings may indicate that both species are attuned to pick up on the same facial features due to their informational value. Put differently, the cognitive mechanisms that process these facial features are shared by both species.

### **Consequences of trait impressions**

In spite of their generally low accuracy, personality trait impressions from faces influence many important decisions (Olivola, Funk, et al., 2014). For example, trustworthiness impressions affect interpersonal trust (van't Wout & Sanfey, 2008), lending decisions (Duarte et al., 2012), and legal sentencing (Porter, ten Brinke, & Gustaw, 2010). A study by Wilson and Rule (2015) even found that untrustworthy-looking criminals were more likely to receive the death penalty (as opposed to life in prison). Moreover, competence impressions influence voting behavior (Olivola & Todorov, 2010a;

Todorov, Mandisodza, Goren, & Hall, 2005) and personnel selection (Graham, Harvey, & Puri, 2017; Ling et al., 2019; Stoker, Garretsen, & Spreuwewers, 2016).

In addition, an extensive literature on the so-called beauty premium shows that facial attractiveness judgments also influence many decisions. Even though impressions of attractiveness are not inaccurate per se (but see Zebrowitz & Rhodes, 2004), they often affect outcomes for which attractiveness should be irrelevant (Maestripieri et al., 2017). In other words, people persistently rely on facial attractiveness to make decisions, which leads to widespread discrimination against unattractive people. For example, individuals with unattractive faces are less likely to be invited for job interviews (Bóo, Rossi, & Urzúa, 2013; Ruffle & Shtudiner, 2015) and receive lower wages (Hamermesh & Biddle, 1994). In general, unattractive people seem to be treated less favorably in social interactions, receiving less help, attention, or other positive outcomes (Langlois et al., 2000; Maestripieri et al., 2017). Thus, how trustworthy, competent, or attractive a person is perceived to be influences a wide range of important social outcomes.

Reliance on split-second judgments of faces is not only widespread, but also surprisingly persistent. People rely on facial appearance even when they have access to objectively better information (Olivola et al., 2018; Rezsescu, Duchaine, Olivola, & Chater, 2012) and even when they are directly told not to (Blair et al., 2004; Hassin & Trope, 2000). Reliance on trait impressions is prevalent among young children and old adults (Charlesworth, Hudson, Cogsdill, Spelke, & Banaji, 2019; Suzuki, 2016) and extends to situations with strong incentives to make unbiased decisions. For example, facial appearance influences the selection and compensation of business leaders even though facial appearance is not related to performance (Graham et al., 2017; Ling et al., 2019). There is also direct evidence that reliance on facial appearance can lead to worse decision outcomes: When judging a variety of characteristics, people are

less accurate if they know what a person looks like, supposedly because they rely in inaccurate facial stereotypes (Olivola & Todorov, 2010b). These studies show that, even if there might be a small kernel of truth in trait impressions, people regularly *overrely* on facial appearance.

### **First impressions as social biases**

Overall, an extensive literature suggests that facial features exert an undue influence on many consequential decisions. This bias not only leads to sub-optimal outcomes for decision-makers, but also to systematic discrimination against people with a certain facial appearance. In other words, people may not only experience unfair treatment because of their gender, race, or sexual orientation, but also because of their facial appearance. However, compared to other types of discrimination, facial discrimination is poorly understood. I therefore argue that we should treat first impressions more like other social biases. Viewing first impressions as biasing influences is not new (Maestripieri et al., 2017; Olivola, Funk, et al., 2014; Zebrowitz et al., 2003). Yet, despite knowledge about the pervasiveness of the bias, many key questions remain unanswered. Approaching first impressions in a similar way as we approach gender or race biases shifts the focus to key questions that need to be addressed if we want to reduce discrimination based on facial appearance.

Specifically, research on social biases is often motivated by three general research questions. How prevalent is the bias? What are the underlying mechanisms? How can we mitigate the bias? We know a lot about the prevalence of facial discrimination, but little about the underlying mechanisms or potential ways to mitigate it. Why do people persistently rely on trait impression, despite their low diagnostic value and despite the fact that this can lead to worse decision outcomes? How can we reduce facial discrimination? This dissertation tries to answer some of these questions.



## Outline of the dissertation

I started this dissertation during an exciting time. Large-scale replications projects revealed that many supposedly robust effects in social psychology (and the behavioral sciences in general) do not withstand closer scrutiny (Camerer et al., 2018; Open Science Collaboration, 2015). Questionable research practices—such as *p*-hacking, hypothesizing after the results are known, and publication bias—appear to be common and drastically increase the rate of false positives in literature (Ioannidis, 2005; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). In light of this crisis of confidence, standards in methodological and statistical practices are changing. Following these recent developments, I took several steps to improve the transparency, reproducibility, and replicability of my work. The majority of studies reported in this dissertation were preregistered. When possible, I conducted a priori power analyses or sensitivity analyses to ensure sufficient statistical power. I also conducted replication studies of my own and others' work. For each chapter, all data, analysis scripts, preregistration documents, and study materials are publically available. Moreover, all papers (including an electronic version of this dissertation) can be found online.

The research presented in this dissertation would not have been possible without the help of my co-authors. The empirical chapters are therefore written using plural personal pronouns. However, because the Introduction and General Discussion also reflect my own thoughts, they are written using singular personal pronouns.

Each empirical chapter in this dissertation is based on an individual paper that is either published or undergoing peer review. The six chapters can be grouped into three sections, corresponding to the three central questions that are raised by a social bias perspective on first impressions. Chapters 2 and 3 replicate and extend previous findings on the consequences of facial appearance. In **Chapter 2** (three studies,  $N =$

## Chapter 1

470), we examine the effect of facial appearance on voting behavior. In a sample of 150 mayoral candidates from 75 constituencies across Italy, we find that attractive-looking (but not competent-looking or trustworthy-looking) politicians receive more votes and are more likely to win. We also test whether trustworthy-looking politicians are more successful in regions with more political corruption, but find no support for this hypothesis.

In **Chapter 3** (one study,  $N = 1,336$ ), we examine the effect of facial appearance on consumer behavior in a popular peer-to-peer market. Data from 1,020 Airbnb listings in New York City shows that attractive-looking (but not trustworthy-looking) hosts charge higher prices for similar apartments, suggesting that people are willing to pay a premium in order to stay in the apartment of an attractive host. Interestingly, we find an attractiveness bias even when people are renting an entire apartment and have little or no direct contact with the host. Together, findings from both chapters support the conclusion that facial appearance influences decision-making in real life.

Chapters 4-6 provide novel insights into the mechanisms underlying reliance on first impressions. These papers were motivated by the observation that biases are often caused by one of two mechanisms (T. D. Wilson & Brekke, 1994). One common source of bias is the existence of false beliefs (i.e., misconceptions). Chapters 4 and 5 explore whether people hold beliefs about the diagnosticity of facial features for judging personality (i.e., physiognomic beliefs). **Chapter 4** (five studies,  $N = 3,861$ ) introduces a scale to measure the prevalence, structure, and correlates of belief in physiognomy. We find that physiognomic beliefs are widespread in the general population. Moreover, individual differences in physiognomic beliefs are associated with overconfidence in first impressions and reliance on first impression in social decision-making.

**Chapter 5** (three studies,  $N = 1,438$ ) replicates and extends the finding that people who believe in physiognomy are more confident in their trait impressions and examines the role of physiognomic beliefs in a more applied setting. Specifically, we find that in a personnel selection context, people who believe in physiognomy view personal photos as more important. They also find it more appropriate and effect to rely on personal photos to make hiring decisions. Results from this chapter also show that people hold heterogeneous beliefs about the manifestation of personality traits in facial features: Sociability is believed to be more reflected in resting faces than morality or competence. This structure is not only reflected in lay beliefs. For example, we find that people are more confident in the accuracy of their sociability judgments. Together, results from Chapters 4 and 5 suggest that prevalent beliefs about the diagnosticity of facial features for inferring personality may explain why people persistently rely on first impressions.

A second common source of bias is automaticity (e.g., misleading intuitions, fluency; T. D. Wilson & Brekke, 1994). Can the quick and efficient processing of faces explain pervasive effects of facial features on social decision-making? This idea is explored in **Chapter 6** (six studies;  $N = 2,732$ ). Results show that reliance on first impressions is relatively effortless and people even prefer to rely on first impressions over another cue that is perceived as more valid, but that is also more effortful to process. Thus, findings from this chapter suggest that people persistently rely on first impressions because they are intuitively accessible, which makes relying on them relatively easy.

In **Chapter 7** (three studies,  $N = 2,274$ ) we attempt to use these new insights to design interventions that can mitigate facial discrimination. First, we create a legal sentencing paradigm that allows us to measure reliance on first impressions at the participant-level. We show that untrustworthy-looking defendants are more likely to be found guilty than trustworthy-looking defendants. Two subsequent studies test the

## Chapter 1

effectiveness of different interventions in reducing reliance on facial trustworthiness (a) by attempting to reduce physiognomic beliefs, or (b) by attempting to disrupt the intuitive accessibility of first impressions. Neither approach successfully reduced facial discrimination. These results underscore the persistence of the bias and highlight the need for future studies on the mitigation of facial discrimination.

In **Chapter 8**, the main findings are summarized and limitations and directions for future research are outlined. Finally, an additional contribution is presented in the **Appendix**. We provide a tutorial on how to use face classification algorithms for detecting a person's gender, age, and race from face images. We also test their accuracy and find that, in many situations, accuracy levels are high and similar to those of human raters. These results suggest that algorithms are a viable alternative to human raters when determining demographic characteristics based face images. Crucially, relying on automated classification procedures can reduce the time spent on data collection. It also allows researchers to test their hypotheses using large, naturalistic data sets (e.g., data from social networking sites, or peer-to-peer markets), as sample size is not constrained by the size of the participant pool.

# Chapter 2

Facial appearance and electoral success:  
Are trustworthy-looking politicians more  
successful in corrupt regions?

Based on:

Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2019). *Facial appearance and electoral success: Are trustworthy-looking politicians more successful in corrupt regions?* Manuscript submitted for publication.

All data, preregistration documents, and analysis scripts are available at the Open Science Framework (<https://osf.io/jdqn2/>).

## **Abstract**

People rely on the facial appearance of political candidates when voting. Here, we examine whether perceptions of competence, trustworthiness, or attractiveness are associated with electoral success in the 2016 Italian local elections. In line with situational leadership theory, we also test whether trait preferences for politicians vary as a function of election context. Specifically, we examine if trustworthy-looking politicians are more successful in regions where political corruption is a salient issue. We analyze electoral data of 150 mayoral candidates from 75 constituencies to test if the association between perceived trustworthiness and electoral success is stronger in Southern Italy, where corruption is more prevalent. Across three preregistered studies ( $N = 470$ ), perceived competence and perceived trustworthiness were not associated with electoral success. Moreover, the influence of trustworthiness perceptions on electoral success did not differ between Southern Italy and the rest of the country. Instead, we found that attractive-looking politicians were more successful. A one standard deviation increase in perceived attractiveness corresponded to a 2.98 percentage point increase in vote share and a 1.91 times increase in the odds of victory. In sum, while our results support the general notion that facial appearance correlates with electoral success, we do not find evidence that corruption moderates the success of trustworthy-looking politicians.

The functioning of democratic political systems requires citizens to elect capable leaders. However, voting decisions are complex and voters often rely on heuristics, simplified decision strategies that require fewer cognitive resources (Quattrone & Tversky, 1988). While some heuristics can lead to accurate inferences under conditions of limited knowledge (e.g., inferring a candidate's stance on policy issues by their party affiliation), other strategies are less justifiable (Kuklinski & Quirk, 2000). For instance, even though trait impressions based on facial appearance are rarely accurate (Olivola & Todorov, 2010b; Todorov & Porter, 2014; but see Lin, Adolphs, & Alvarez, 2018), they predict the electoral success of political candidates (Olivola & Todorov, 2010). In other words, people rely on appearance-based trait impressions when voting.

A host of studies has shown that competent-looking politicians enjoy more political success (Antonakis & Dalgas, 2009; Ballew & Todorov, 2007; Todorov et al., 2005). However, some results suggest that voters are also influenced by other apparent traits, such as the perceived attractiveness, dominance, or sociability of candidates (Berggren, Jordahl, & Poutvaara, 2010; Castelli, Carraro, Ghitti, & Pastore, 2009; F. F. Chen, Jing, & Lee, 2014; Jäckle, Metz, Wenzelburger, & König, 2019). This raises the question whether effects of specific traits systematically vary across different contexts. Evidence from controlled lab experiments provide initial support for this notion. For example, Little and colleagues (2012) showed that framing a hypothetical election as taking place during a time of war or peace influenced participants' preference for trustworthy-looking or attractive-looking candidates. Specifically, attractive-looking leaders were favored more strongly during a time of war, while trustworthy-looking leaders were favored more strongly during a time of peace, suggesting that which (apparent) traits are favored in politicians may be influenced by the political context in which an election is taking place. However, it is thus far unclear whether the moderating role of election context generalizes to other

contextual frames, and whether this effect influences real-world elections.

Here, we analyze results of the 2016 Italian local elections to examine the influence of facial appearance on voting behavior. First, we test whether the perceived attractiveness, competence, or trustworthiness of candidates is related to their electoral success. Second, we investigate whether the salience of a political issue—the regional prevalence of institutional corruption—moderates the association between trait perceptions and electoral success. Specifically, we hypothesize that voters are more motivated to elect a trustworthy leader when corruption is a salient issue. As a consequence, trustworthy-looking candidates should be more successful in regions where corruption is more prevalent (e.g., in Southern Italy vs. the rest of the country; Linhartova & Pultarova, 2015).

### **Election context moderates trait preferences**

When asked directly which personality traits a politician should possess, voters primarily mention competence (Miller, Wattenberg, & Malanchuk, 1986; Sussman, Petkova, & Todorov, 2013). In line with this explicit preference, Todorov and colleagues (2005) found that appearance-based impressions of competence, but not impressions of trustworthiness, likeability, or attractiveness, are associated with success in elections for the US Senate and House of Representatives. The notion that voters rely on the facial appearance of candidates to make voting decisions was supported by many subsequent studies, which investigated the relationship between facial appearance and political success in a wide range of countries and electoral systems (for a review, see Olivola & Todorov, 2010a). While most studies found that competent-looking politicians are more successful (Antonakis & Dalgas, 2009; Ballew & Todorov, 2007; Castelli et al., 2009; Sussman et al., 2013), impressions of other traits also predict electoral success under certain



conditions. For example, Berggren and colleagues (2010) found a positive effect of attractiveness on voting behavior in Finland (for similar results in the United States, see Jäckle et al., 2019). In other studies, electoral success was related to perceptions of dominance (F. F. Chen et al., 2014; Sussman et al., 2013), sociability (Castelli et al., 2009), or gender-typicality (Hehman, Carpinella, Johnson, Leitner, & Freeman, 2014).

To account for these findings, researchers have started investigating how the context in which an election is taking place influences the association between specific trait impressions and electoral success. For example, cross-cultural data suggests that competence-related traits are more predictive of electoral success in Western societies (e.g., the United States) than in East Asian societies (e.g., Japan or Taiwan), whereas the opposite pattern holds for trustworthiness-related traits (F. F. Chen, Jing, Lee, & Bai, 2016; Rule et al., 2010). Next to cultural differences, trait preferences may also vary as a function of the political context in which an election is taking place. In general, voters may prefer different traits in political leaders depending on which political issues are particularly salient. This idea follows from situational leadership theory, which stress that leader selection is context-sensitive, with different leader types being favored depending on which tasks they are expected to perform (Epitropaki & Martin, 2004; Hollander & Julian, 1969; Olivola, Eubanks, & Lovelace, 2014; Yukl, 1989).

Following this reasoning, Little and colleagues (2014; 2007; 2012) demonstrated that participants' hypothetical voting behavior can be influenced by manipulating the political context of an election. They found that participants had a stronger preference for individuals whose facial trustworthiness had been digitally enhanced in a time of peace, whereas individuals whose facial attractiveness had been digitally enhanced were more strongly favored in a time of war. This pattern

suggests that prosocial traits (i.e., trustworthiness) are favored in leaders when the political context is characterized by collaboration, but traits related to health and formidability (i.e., attractiveness) are favored when the political climate is characterized by conflict (for similar results, see Ferguson et al., 2019; Laustsen & Petersen, 2015; Spisak, Dekker, Krüger, & Van Vugt, 2012).

The influence of war vs. peace frames on hypothetical votes provides initial evidence for the context-sensitive nature of face-based leader choice in the political domain. It also suggests that impressions of trustworthiness—a trait which should be highly desirable in a politician (Miller et al., 1986)—may influence voting behavior under some (but not all) conditions. It is unclear though (a) whether the effect of political context generalizes to issues other than a country's state of war or peace and (b) whether it extends to real-world elections.

Here, we examine if the salience of a different political issue—the prevalence of institutional corruption—moderates which trait perceptions predict electoral success. Corruption is a recurring issue for political systems around the world and considerable resources are devoted to monitoring and diminishing corrupt practices (Jain, 2001). It is also a salient issue for voters as corruption charges lead to a substantial loss in votes (J. G. Peters & Welch, 1980; Welch & Hibbing, 1997). Building on these observations, we hypothesize that voters are more motivated to elect a trustworthy candidate when corruption is a salient issue. As a consequence, trustworthy-looking politicians should be more successful in constituencies with high levels of institutional corruption.

### **Studies 2.1-2.3**

In the current investigation, we attempt to replicate the finding that voters rely on trait impressions from faces when deciding whom to elect. Specifically, we test which traits are associated with of electoral success

and whether the effect of trait impressions on voting varies as a function of political context. To this end, we examine the effect of facial appearance on electoral success in the 2016 Italian local elections. We focus on Italy because Italy exhibits large regional differences in the prevalence of corruption, with substantially higher levels in the south (Del Monte & Papagni, 2007; Fiorino, Galli, & Petrarca, 2012; Linhartova & Pultarova, 2015). We therefore test whether trustworthy-looking politicians are more successful in Southern Italy (compared to the rest of the country).

In all three studies, we measured the perceived competence, trustworthiness, and attractiveness of 150 mayoral candidates from 75 constituencies where two-candidate elections took place. We used ratings on the three trait dimensions to predict (a) the winner of the election and (b) the margin of victory. Results can differ substantially due to small differences in the methodology of studies (Landy et al., in press). To probe the robustness of our results, we therefore varied whether trait impressions were obtained from American (Study 2.1) or Dutch participants (Studies 2.2 and 2.3).<sup>3</sup> We also varied whether trait perceptions were assessed with a single trait item (e.g., ratings of trustworthiness; Studies 2.1 and 2.3) or with multiple trait items (e.g., ratings of trustworthiness, honesty, and fairness for measuring trustworthiness; Study 2.2). Finally, we varied whether participants indicated their trait impressions by selecting the candidate scoring higher on a given trait dimension in a two-alternative forced-choice

---

<sup>3</sup> We recruited participants from outside of Italy to ensure that most, if not all, participants are unfamiliar with the political candidates. In general, there is substantial cross-cultural agreement in trait ratings from faces (Cunningham, Roberts, Barbee, & Druen, 1995; Rule et al., 2010). Moreover, previous studies have shown that, for example, trait ratings by American participants predict election outcomes in Bulgaria (Sussman et al., 2013) and trait ratings by Germany participants prediction election outcomes in the United States (Jäckle et al., 2019).

format (Study 2.1) or by rating all candidates sequentially on a continuous scale (Studies 2.2 and 2.3).

## Methods

The studies were preregistered and all data, preregistration documents, and analysis scripts are available at the Open Science Framework (<https://osf.io/jdqn2/>).<sup>4</sup> We report how our sample sizes were determined, all data exclusions, and all measures in the studies.

**Participants.** For each trait dimension, ratings from at least 29 independent raters were collected, as previous studies have shown that this provides relatively stable average ratings (Hehman, Xie, Ofosu, & Nespoli, 2018). We asked participants at the end of each study whether they had recognized any of the individuals that were shown in the photos and, in case they answered affirmatively, who they had recognized. While some participants claimed to have recognized at least one candidate (Study 2.1: 6.12%, Study 2.2: 8.64%, Study 2.3: 14.29%), none provided correct names or mentioned the fact that the depicted individuals are Italian politicians.<sup>5</sup>

**Study 2.1.** Participants were 160 workers from Amazon Mechanical Turk (Paolacci & Chandler, 2014) who completed the study in return for \$1. Thirteen participants (8.13%) who failed an attention check at the end of the study were excluded, leaving a final sample of 147 participants (43.54% female;  $M_{\text{age}} = 32.07$ ,  $SD_{\text{age}} = 8.25$ ). On average,

---

<sup>4</sup> For Study 2.1 and Study 2.2, we preregistered to conduct multilevel regression analyses to account for the fact that individual candidates are nested within different municipalities. However, this analysis is not suitable given the dyadic structure of the data in which one candidate's vote share and election outcome is perfectly mirrored by the other candidate's vote share and election outcome. We therefore chose to follow a different analysis strategy to account for the dependencies in our data (see the Results section for more details) and conducted a third study for which the correct analyses were specified a priori.

<sup>5</sup> Among the people that were purportedly recognized were the actor James Franco, the boxer Nasseem Hamad, and a participant's dentist.

candidates were rated by 47 participants ( $Min = 43$ ,  $Max = 54$ ) on each of the three trait dimension (competence, trustworthiness, and attractiveness).

**Study 2.2.** Participants were 223 Dutch undergraduate psychology students from Tilburg University who participated in return for partial course credit. Three participants (1.35%) who provided the same response across all trials were excluded, leaving a final sample of 231 participants (76.92% female;  $M_{age} = 20.18$ ,  $SD_{age} = 2.31$ ). The final sample size was based on the number of students that participated in the study within two weeks. On average, candidates were rated by 31 participants ( $Min = 29$ ,  $Max = 33$ ) on each of the three trait dimension.

**Study 2.3.** Participants were 93 Dutch undergraduate psychology students from Tilburg University who participated in return for partial course credit. One participant (1.08%) who provided the same response across all trials was excluded, leaving a final sample of 92 participants (49.45% female;  $M_{age} = 20.87$ ,  $SD_{age} = 2.22$ ). The final sample size was based on the number of students that participated in the study within two weeks. On average, candidates were rated by 30 participants ( $Min = 30$ ,  $Max = 31$ ) on each trait dimension.

**Materials.** We retrieved the results of the 2016 Italian local elections. Residents of cities with a population greater than 15,000 could directly vote for different mayoral candidates in a multi-candidate two-round system. For the current analysis, we focused on 126 constituencies in which no candidate received the absolute majority in the first round. In that case, the two candidates who received the most votes competed in a second round which was held two weeks later. In line with our preregistered exclusion criteria, elections with at least one female candidate (36 elections, 28.57%) were excluded to remove the confounding role of gender (Chiao, Bowman, & Gill, 2008).

Next, images of the candidates were downloaded from the internet. We selected photos in which candidates faced the camera with their

faces being completely visible. For most candidates, we selected the photo from their election poster, as this was the photo that most voters were exposed to prior to the election. If the election poster could not be retrieved, another photo was selected. The wide majority of candidates were smiling in their photos and we tried to ensure that differences in affective expression between the two candidates were minimal. If one candidate showed a broad smile while the other looked neutral and no other photos could be found for the latter, then the election was excluded from analysis (15 elections, 16.67%). This resulted in a final sample of 75 elections with a total of 150 candidates.

The photos were converted to grayscale, cropped so that only the candidate's face and hair were visible, and resized to a height of 300 pixels. For each election, we recorded which of the two candidates won and their margin of victory, which constituted our dependent variables. We also recorded and whether the candidate was the incumbent or running against the incumbent and whether the constituency was located in the south ( $n = 34$ ). Southern Italy encompasses the administrative regions of Abruzzo, Apulia, Basilicata, Campania, Calabria, Molise, and Sicily.

**Procedure.** To measure candidates' perceived competence, trustworthiness, and attractiveness, participants, who were unaware of the context of the study and the identity of the people shown in the photos, evaluated all candidates on one specific trait dimensions. Each participant rated the candidates on only one trait in order to avoid consistency effects in ratings (Penton-Voak, Pound, Little, & Perrett, 2006).

**Study 2.1.** In Study 2.1, binary trait ratings on three dimensions were collected. Participants were randomly allocated to one of three conditions which determined whether they would rate the candidates' competence, trustworthiness, or attractiveness. They saw the 75 pairs of candidates in a random order and were asked to select the candidate that

looks more *competent*, *trustworthy*, or *attractive* depending on the condition. The percentage of participants who selected a given candidate as scoring higher than his opponent served as our measure of perceived competence, trustworthiness, and attractiveness.

**Study 2.2.** In Study 2.2, trait ratings were assessed with Likert scales. Participants were randomly allocated to one of the seven trait conditions: Perceived competence was measured via ratings of *competence*, *capability*, and *intelligence*, perceived trustworthiness via ratings of *trustworthiness*, *honesty*, and *fairness*, and perceived attractiveness via ratings of *attractiveness*. The 150 photos were rated in a random order on a 9-point scale ranging from *not at all* [trait] (1) to *extremely* [trait] (9). We calculated the intraclass correlation coefficients (ICC) as a measure of inter-rater consistency (Shrout & Fleiss, 1979). Participants showed significant consensus in their ratings (competence:  $ICC(2, 1) = .089$ , capability:  $ICC(2, 1) = .110$ , intelligence:  $ICC(2, 1) = .176$ , trustworthiness:  $ICC(2, 1) = .106$ , honesty:  $ICC(2, 1) = .111$ , fairness:  $ICC(2, 1) = .151$ , attractiveness:  $ICC(2, 1) = .331$ ; all  $ps < .001$ ).

A confirmatory factor analysis indicated that a three-factor structure adequately fit ratings on the seven trait dimensions:  $\chi^2 = 882.54$  ( $df = 21$ ), RMSEA = 0.87, SRMR = 0.36, CFI = 0.98. Therefore, ratings of *competence*, *capability*, and *intelligence* were averaged to form a competence score; ratings of *trustworthiness*, *honesty*, and *fairness* were averaged to form a trustworthiness score; and ratings of attractiveness constituted a candidate's attractiveness score. We created relative trait scores per election by subtracting the runner-up's trait score from the winner's trait score. In other words, each candidate's trait scores reflected their perceived trustworthiness, competence, or attractiveness *relative* to their opponent.

**Study 2.3.** In Study 2.3, trait ratings were assessed with Likert scales. Participants were randomly allocated to one of three conditions which determined whether they would rate candidates' *competence*,

*trustworthiness*, or *attractiveness*. The 150 images were rated in a random order on a 9-point scale ranging from *not at all* [trait] (1) to *extremely* [trait] (9). We calculated the intraclass correlation coefficient as a measure of inter-rater consistency. Participants showed significant consensus in their ratings (competence:  $ICC(2, 1) = .113$ , trustworthiness:  $ICC(2, 1) = .123$ , attractiveness:  $ICC(2, 1) = .222$ ; all  $ps < .001$ ). For each candidate, ratings on the three dimensions were averaged across all participants and this served as our measure of perceived competence, trustworthiness, and attractiveness. We again created relative trait scores per election by subtracting the runner-up's trait score from the winner's trait score.

**Analysis plan & sensitivity analysis.** All trait scores were z-standardized to allow for comparisons between the studies. In all three studies, we tested for the effects of facial appearance on election outcomes by predicting in separate models (a) the winner of the election and (b) the margin of victory with candidates' perceived facial competence, trustworthiness, and attractiveness. We also tested whether the effect of trait perceptions varied as a function of the geographical location of a constituency (south vs. rest of the country). We control for incumbency status of the two candidates in all regression analyses as incumbents often have the advantage over political challengers (G. W. Cox & Katz, 1996). All analyses were conducted in R (R Core Team, 2019).

For each effect of interest, a sensitivity analysis was conducted to determine the minimum effect size we were able to detect with 80% power (and  $\alpha = .05$ ). As software commonly used for sensitivity analyses, such as G\*Power (Faul, Erdfelder, Lang, & Buchner, 2007), does not support dyadic data, we relied on the *simr* package in R (Green & Macleod, 2016). The package does not provide a function specifically designed for conducting sensitivity analyses. However, it can provide estimates of observed power for coefficients in regression models. For



each of our models, we systematically varied the effect size for all effects of interest and computed observed power. Performing power calculations across a range of effect sizes allowed us to determine the minimum effect size at which our model had at least 80% power to detect a significant effect.

Regarding the effects of perceived competence, trustworthiness, and attractiveness on the percentage of received votes (i.e., the winner's margin of victory), analyses showed that we had 80% power to detect an increase of 2.10 percentage points, 2.19 percentage points, and 2.08 percentage points, respectively. We had 80% power to detect a difference of 2.14 percentage points for the interaction effect between perceived trustworthiness and geographical location of the constituency. Regarding the effects of perceived competence, trustworthiness, and attractiveness on the likelihood of a candidate's success, analyses showed that we had 80% power to detect odds ratios of 1.60, 1.69, and 1.62, respectively. We had 80% power to detect an odds ratio of 1.59 for the interaction effect between perceived trustworthiness and geographical location of the constituency.

## Results

The average margin of victory was 7.84 percentage points with a median of 5.72 percentage points ( $SD = 6.54$ ,  $Min = 0.14$ ,  $Max = 25.05$ ). Twenty-five elections (33%) featured an incumbent. For each study, we computed correlations between candidates' perceived attractiveness, competence, and trustworthiness and incumbency status. A random-effects meta-analysis across the three studies showed that incumbency status was positively correlated with competence ratings,  $r = .10$ ,  $p = .029$ . There were no significant correlation between incumbency status and attractiveness ratings,  $r = -.07$ ,  $p = .12$ , or trustworthiness ratings,  $r = .06$ ,  $p = .21$ . Trustworthiness ratings were moderately correlated with competence ratings,  $r = .42$ ,  $p < .001$ , and attractiveness ratings,  $r = .47$ ,

$p < .001$ . There was no significant correlation between competence ratings and attractiveness ratings,  $r = .07$ ,  $p = .33$ .

We also conducted an exploratory analysis of Google Trends data to test if political corruption is a more salient issue in Southern Italy. Google Trends provides access to the number of search queries for specific terms across different time frames and geographical locations (Choi & Varian, 2012). We recorded the number of searches that contained the word “corruption” (in Italian) across different Italian regions for four time windows: one month, three months, six months, and twelve months prior to the election. For each time window, the number of searches were rescaled to range from 0 to 100. Data for two southern regions was unavailable for the one-month time window.

Corruption-related searches were more prevalent in southern regions (vs. the rest of the country) one month prior to the election (south:  $M = 42.36$ ,  $SD = 23.65$ , rest:  $M = 24.54$ ,  $SD = 9.47$ ),  $t(24.67) = 3.39$ ,  $p = .002$ ,  $d = 0.90$ , three months prior to the election (south:  $M = 58.18$ ,  $SD = 14.48$ , rest:  $M = 41.59$ ,  $SD = 14.24$ ),  $t(70.01) = 4.98$ ,  $p < .001$ ,  $d = 1.15$ , six months prior to the election (south:  $M = 64.50$ ,  $SD = 13.04$ , rest:  $M = 53.85$ ,  $SD = 15.37$ ),  $t(72.95) = 3.25$ ,  $p = .002$ ,  $d = 0.75$ , and twelve months prior to the election (south:  $M = 64.82$ ,  $SD = 14.86$ , rest:  $M = 56.27$ ,  $SD = 14.10$ ),  $t(68.94) = 2.54$ ,  $p = .013$ ,  $d = 0.59$ . These results lend support to our assumption that corruption is a more salient issue in the south of Italy.

**Margin of victory.** First, we examined whether candidates’ facial appearance predicted the margin of victory (i.e., candidates’ relative vote share). We estimated OLS regression models in which vote share was simultaneously regressed on candidates’ perceived attractiveness, competence, and trustworthiness. Due to the dyadic structure of our data, for any given election, one candidate’s data (e.g., their vote share, their relative attractiveness) always perfectly mirrored their opponent’s data. To account for this dependency, we randomly selected one

candidate from each election and conducted our analyses on this sample of 75 candidates. However, results of these analyses vary depending on the specific combination of winners and runner-ups that are sampled. Therefore, we selected and analyzed 100,000 random samples of 75 candidates that included one candidate from each election. We calculated mean estimates for our predictors across all randomly drawn samples. This bootstrapping procedure was performed for each study. Finally, the results of the three studies were aggregated in a random-effects meta-analysis (see Figure 1).<sup>6</sup>

Results showed that, across the three studies, perceived attractiveness positively predicted vote share,  $\beta = 2.977$ ,  $SE = 0.765$ ,  $z = 3.89$ ,  $p < .001$ . Surprisingly, perceived competence *negatively* predicted vote share,  $\beta = -1.676$ ,  $SE = 0.745$ ,  $z = 2.25$ ,  $p = .025$ . Candidates who scored one standard deviation higher on attractiveness received 2.98 percentage points more votes whereas candidates who scored one standard deviation higher on competence received 1.68 percentage points fewer votes. We did not find evidence that perceived trustworthiness was related to vote share,  $\beta = 0.102$ ,  $SE = 0.818$ ,  $z = 0.12$ ,  $p = .90$ .<sup>7</sup>

We also examined the influence of regional variation in corruption by estimating a second model in which we added an interaction effect between the geographical region in which an election took place (dummy-coded 0.5 for the south and -0.5 for all other regions) and candidates' perceived trustworthiness. This interaction effect was significant but in the opposite direction of our hypothesis (i.e.,

---

<sup>6</sup> The pattern of results did not change when analyzing median estimates, instead of mean estimates. Detailed statistics for each study can be found in the Supplemental Materials

<sup>7</sup> We also regressed vote share on each trait dimension separately. This again yielded a positive effect of perceived attractiveness,  $\beta = 2.823$ ,  $SE = 0.681$ ,  $z = 4.15$ ,  $p < .001$ , and no effect of perceived trustworthiness,  $\beta = 0.751$ ,  $SE = 0.697$ ,  $z = 1.08$ ,  $p = .28$ . The effect of perceived competence was negative but only marginally significant,  $\beta = -1.295$ ,  $SE = 0.703$ ,  $z = 1.84$ ,  $p = .066$ .

## Chapter 2

trustworthiness had *less* of an impact in the south),  $\beta = -2.907$ ,  $SE = 1.379$ ,  $z = 2.12$ ,  $p = .034$ . Perceived trustworthiness was negatively associated with vote share in the south,  $\beta = -3.851$ ,  $SE = 1.281$ ,  $z = 3.01$ ,  $p = .003$ , and positively associated with vote share in the north,  $\beta = 3.114$ ,  $SE = 1.048$ ,  $z = 2.97$ ,  $p = .003$ . We also explored whether associations between vote share and perceived attractiveness or competence differed between Southern Italy and the rest of the country. There was no significant interaction effect between region and perceived attractiveness,  $\beta = 0.190$ ,  $SE = 1.412$ ,  $z = 0.13$ ,  $p = .89$ , but the interaction effect between region and perceived competence was significant,  $\beta = 4.361$ ,  $SE = 1.359$ ,  $z = 3.21$ ,  $p = .001$ . Perceived competence was positively associated with vote share in the south,  $\beta = 2.832$ ,  $SE = 1.406$ ,  $z = 2.01$ ,  $p = .044$ , but negatively associated with vote share in the north,  $\beta = -4.356$ ,  $SE = 0.851$ ,  $z = 5.12$ ,  $p < .001$ .

In sum, candidates' perceived attractiveness and competence, but not their perceived trustworthiness, predicted their vote share with attractive-looking candidates receiving more votes and competent-looking candidates receiving fewer votes. The association between perceived trustworthiness and vote share significantly differed between Southern Italy and the rest of the country. However, the observed pattern was opposite to our prediction: More trustworthy-looking politicians received fewer votes in the south, but more votes in the north.

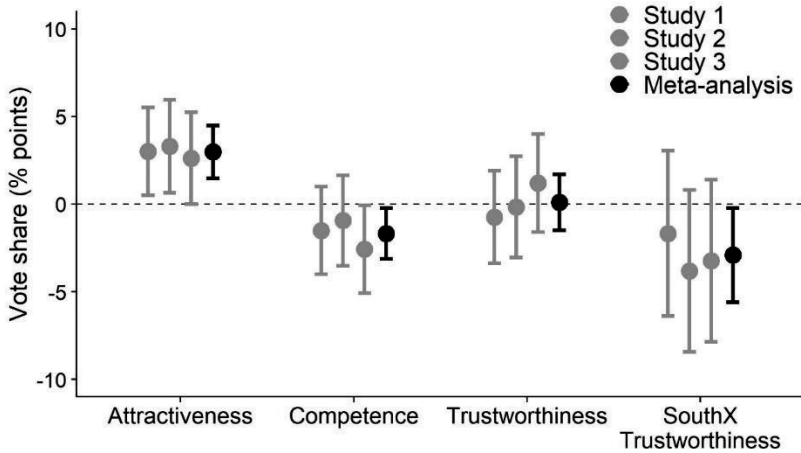


Figure 2.1. The influence of winners' facial appearance on the margin of victory. The graph displays the results of the three studies (starting with Study 2.1 on the left) and the meta-analytic estimates. In a first model, vote share was regressed on candidates' perceived trustworthiness, competence, and attractiveness while controlling for incumbency status. In a second model, an interaction term between the region in which an election took place (dummy-coded 0.5 for the south and -0.5 for all other regions) and perceived trustworthiness was added.

**Electoral success.** Next, we examined whether candidates' facial appearance predicted their likelihood of winning the election. We estimated logistic regression models in which election outcome (0 = candidate lost, 1 = candidate won) was regressed on the candidates' perceived attractiveness, competence, and trustworthiness. We followed the same bootstrapping procedure as described before and the results of the three studies were again aggregated in a random-effects meta-analysis (see Figure 2).

Results showed that, across the three studies, perceived attractiveness predicted electoral success,  $\beta = 0.647$ ,  $SE = 0.179$ ,  $z = 3.62$ ,  $p < .001$ . Candidates who scored one standard deviation higher on attractiveness were 1.91 times more likely to win their election. We did not find evidence that perceived competence,  $\beta = -0.260$ ,  $SE = 0.162$ ,  $z =$

## Chapter 2

1.61,  $p = .11$ , or perceived trustworthiness were related to electoral success,  $\beta = -0.100$ ,  $SE = 0.178$ ,  $z = 0.56$ ,  $p = .58$ .<sup>8</sup>

We also examined the influence of regional variation in corruption on the predictive power of perceived trustworthiness. A second model was estimated in which we added an interaction effect between the geographical region in which an election took place (dummy-coded 0.5 for the south and -0.5 for all other regions) and candidates' perceived trustworthiness. This interaction effect was not significant,  $\beta = -0.505$ ,  $SE = 0.310$ ,  $z = 1.63$ ,  $p = .10$ . We also explored whether associations between electoral success and perceived attractiveness or competence differed between Southern Italy and the rest of the country. There were no significant interaction effects between region and perceived attractiveness,  $\beta = -0.140$ ,  $SE = 0.331$ ,  $z = 0.42$ ,  $p = .67$ , or region and perceived competence,  $\beta = 0.450$ ,  $SE = 0.317$ ,  $z = 1.41$ ,  $p = .16$ .

In sum, candidates' perceived attractiveness, but not their perceived competence or trustworthiness, predicted their likelihood of winning the election with attractive-looking candidates being more successful. Crucially, we did not find evidence that perceived trustworthiness is more predictive of electoral success in Southern Italy compared to the rest of the country.

---

<sup>8</sup> We also regressed election outcome on each trait dimension separately. This again yielded a positive effect of perceived attractiveness,  $\beta = 0.544$ ,  $SE = 0.154$ ,  $z = 3.52$ ,  $p < .001$ , no effect of perceived competence,  $\beta = -0.223$ ,  $SE = 0.142$ ,  $z = 1.57$ ,  $p = .12$ , and no effect of perceived trustworthiness,  $\beta = 0.076$ ,  $SE = 0.138$ ,  $z = 0.55$ ,  $p = .58$ .

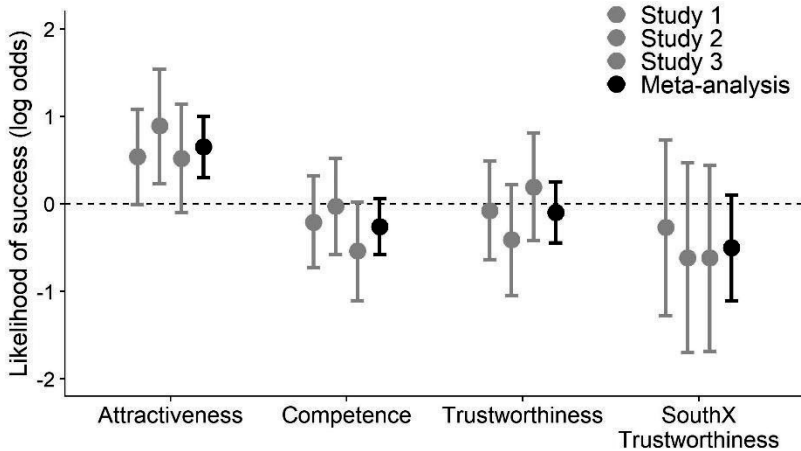


Figure 2.2. The influence of candidates' facial appearance on their likelihood of winning the election. The graph displays the results of the three studies (starting with Study 2.1 on the left) and the meta-analytic estimates. In a first model, election outcome was regressed on candidates' perceived trustworthiness, competence, and attractiveness while controlling for incumbency status. In a second model, an interaction term between the geographical region in which an election took place (dummy-coded 0.5 for the south and -0.5 for all other regions) and perceived trustworthiness was added.

**Robustness checks.** We conducted several exploratory analyses to probe the robustness of our results. First, we re-ran our regression models with additional control variables. We included population size, voter turnout, and dummy variables indicating whether the candidate was a member of the Five Star Movement or running against one. The Five Star Movement is a recently established party whose political agenda includes a strong anti-establishment and anti-corruption stance (Mosca, 2014). Perceived attractiveness still positively predicted vote share,  $\beta = 2.678$ ,  $SE = 0.769$ ,  $z = 3.48$ ,  $p < .001$ , and there was no effect of perceived trustworthiness,  $\beta = -0.609$ ,  $SE = 0.809$ ,  $z = 0.75$ ,  $p = .45$ . The effect of perceived competence was no longer significant,  $\beta = -0.928$ ,  $SE = 0.754$ ,  $z = 1.23$ ,  $p = .22$ . There was no significant interaction effect between region and perceived trustworthiness on vote share,  $\beta = 0.001$ ,

## Chapter 2

$SE = 3.248$ ,  $z < 0.01$ ,  $p > .99$ . Moreover, perceived attractiveness still positively predicted electoral success,  $\beta = 0.599$ ,  $SE = 0.191$ ,  $z = 3.14$ ,  $p = .002$ , but perceived competence,  $\beta = -0.181$ ,  $SE = 0.180$ ,  $z = 1.01$ ,  $p = .31$ , and perceived trustworthiness,  $\beta = -0.122$ ,  $SE = 0.193$ ,  $z = 0.63$ ,  $p = .53$ , did not. There was also no significant interaction effect between region and perceived trustworthiness on electoral success,  $\beta = 0.007$ ,  $SE = 0.757$ ,  $z = 0.01$ ,  $p = .99$ .

Second, our analyses included two municipalities located on the island of Sardinia. While Sardinia is sometimes treated as a separate region altogether (e.g., Bigoni, Bortolotti, Casari, Gambetta, & Pancotto, 2016), it is similar to the island of Sicily and the southern part of mainland Italy—both of which were coded as being part of the south—in regards to the prevalence of corruption (Fiorino et al., 2012). We therefore recoded the two Sardinian municipalities as belonging to the south (i.e., the region where we expected to find stronger effects of perceived trustworthiness). The interaction effect between region and perceived trustworthiness was only marginally significant for predicting the vote share,  $\beta = -2.653$ ,  $SE = 1.382$ ,  $z = 1.92$ ,  $p = .055$ , and not significant for predicting the likelihood of success,  $\beta = -0.314$ ,  $SE = 0.308$ ,  $z = 1.02$ ,  $p = .31$ .

Finally, we analyzed Google Trends data for a more fine-grained analysis of how regional variation in corruption salience influence the success of trustworthy-looking candidates. Thus, instead of including a dummy variable for region (south vs. rest of the country), we included a variable indicating the relative frequency of Google searches that included the word “corruption” in the specific region (ranging from 0 indicating no searches to 100 indicating the number of searches in the region with the highest search frequency). We analyzed four time windows: one month, three months, six months, and twelve months prior to the election. We found no significant interaction effect between the number of Google searches and perceived trustworthiness on vote



share (one-month window:  $\beta = -0.047$ ,  $SE = 0.065$ ,  $z = 0.72$ ,  $p = .47$ , three-month window:  $\beta = -0.036$ ,  $SE = 0.054$ ,  $z = 0.68$ ,  $p = .50$ , six-month window:  $\beta = 0.025$ ,  $SE = 0.057$ ,  $z = 0.44$ ,  $p = .66$ , twelve-month window:  $\beta = 0.004$ ,  $SE = 0.054$ ,  $z = 0.07$ ,  $p = .94$ ) or electoral success (one-month window:  $\beta = 0.009$ ,  $SE = 0.016$ ,  $z = 0.56$ ,  $p = .58$ , three-month window:  $\beta = -0.005$ ,  $SE = 0.012$ ,  $z = 0.45$ ,  $p = .66$ , six-month window:  $\beta = 0.009$ ,  $SE = 0.012$ ,  $z = 0.68$ ,  $p = .49$ , twelve-month window:  $\beta = -0.009$ ,  $SE = 0.012$ ,  $z = 0.73$ ,  $p = .47$ ).

## General discussion

The current set of studies investigated the association between appearance-based trait impressions and voting decisions. Analyzing data from the 2016 Italian local elections, we found that more attractive-looking candidates received more votes and were more likely to win their election. The size of this relationship was not trivial. Candidates who scored one standard deviation higher on perceived attractiveness received 2.98 percentage points more votes and were 1.91 times more likely to win. To put the increase in vote share in perspective, 22 of the 75 elections we studied here were decided by fewer than 3 percentage points. We found no evidence that the perceived trustworthiness of candidates was related to their electoral success. Our results are therefore in line with previous studies showing that attractive politicians are more successful (Berggren et al., 2010; Jäckle et al., 2019; King & Leigh, 2009; Rosar, Klein, & Beckers, 2008), whereas perceived trustworthiness seems to be unrelated to electoral success (Berggren et al., 2010; Todorov et al., 2005).

The results for perceived competence were more ambiguous. Contrary to previous studies (Ballew & Todorov, 2007; Todorov et al., 2005), we found a negative effect of perceived competence on vote share. However, the effect was smaller than the minimum effect size we were able to detect with 80% power, making a false positive result more likely.

## Chapter 2

Exploratory analyses also showed that the effect was not robust to controlling for additional variables (e.g., turnout) and we found no significant effect of competence on the likelihood of success. We therefore conclude that the present results do not provide clear support for the notion that perceived competence is positively or negatively related to electoral success.

While our findings replicate the effect of facial appearance on voting behavior, our main aim was to test predictions from situational leadership theory, which emphasizes the context-specific nature of leader selection (Epitropaki & Martin, 2004; Hollander & Julian, 1969; Olivola, Eubanks, et al., 2014; Yukl, 1989). To this end, we examined whether the effect of trait perceptions were moderated by the political context in which the elections took place. Framing an election as taking place during a time of war vs. a time of peace has been shown to influence participants' preferences for attractive-looking and trustworthy-looking candidates in a simulated voting environment (Little et al., 2012). Here, we tested if a different political issue, the prevalence of institutional corruption, influences voters' trait preferences in real life elections. We reasoned that trustworthy-looking politicians would be more successful Southern Italy where corruption is more prevalent than in the rest of the country and voters may be more motivated to elect a trustworthy leader.

Across three studies, we found no evidence that trustworthy-looking candidates were more successful in Southern Italy. Some analyses even revealed the opposite pattern: Perceived trustworthiness was negatively related to vote share in Southern Italy, but positively related to vote share in the rest of the country. However, this pattern of results did not emerge when including additional control variables or when analyzing candidates' likelihood of winning an election. We also analyzed the frequency of Google searches that included the term corruption as a proxy for the salience of political corruption. Search queries were more frequent in the south, suggesting that corruption is

indeed a more salient issue in Southern Italy, but the salience of corruption (as measured with Google searches) did not moderate the electoral success of trustworthy-looking politicians. Overall, the present results do not provide support for the hypothesis that trustworthy-looking candidates are more successful in regions where political corruption is more salient.

In fact, we did not find any evidence that apparent trustworthiness positively affected vote share or the likelihood of electoral success. This may seem surprising as morality judgments are a strong predictor of overall person evaluations (Brambilla, Rusconi, Sacchi, & Cherubini, 2011; Goodwin, Piazza, & Rozin, 2014) and people name morality-related traits such as honesty and incorruptibility as traits that a politician should possess (Miller et al., 1986; Sussman et al., 2013). Previously reported null results for the effect of perceived trustworthiness on electoral success may have been due to the fact that the majority of studies were conducted in countries with relatively low levels of institutional corruption, such as Finland (Berggren et al., 2010) and the United States (Todorov et al., 2005). In these countries, voters may be less concerned with electing a potentially corrupt leader and place more weight on other traits such as competence. However, investigations in countries with higher levels of corruption such as Bulgaria (Sussman et al., 2013) or Italy (Castelli et al., 2009) did not find a positive relationship between trustworthiness-related traits (e.g., morality, honesty/incorruptibility) and electoral success either. The current results are in line with these findings and suggest that the apparent trustworthiness of political candidates is not associated with their electoral success, even in countries where levels of institutional corruption are relatively high.

### **Limitations and future directions**

Given the higher levels of corruption in Southern Italy, some candidates in our sample might have been accused of (or even directly involved in) corrupt practices. Even if voters in the south are more concerned with electing a trustworthy candidate, corruption allegations probably constitute a stronger indicator of a candidate's trustworthiness, overriding any effect of facial appearance. In other words, given the prevalence of corruption in Italy, voters might be more knowledgeable about the trustworthiness of candidates because of their (alleged) link to corrupt practices and rely on this knowledge, rather than appearance-base impressions, when making voting decision. Future studies could circumvent this issues by investigating the effect of corruption under more controlled conditions in the lab. Following the procedure of Little and colleagues (2012), it could be tested whether participants vote more often for individuals whose perceived trustworthiness was digitally enhanced when institutional corruption is made salient.

An alternative interpretation of the current null results is that trait preferences for politicians are in fact stable and that the same trait perceptions affect voting decisions across different contexts. However, this view cannot explain why different traits have been linked to electoral success in previous studies (Berggren et al., 2010; Castelli et al., 2009; F. F. Chen et al., 2014; Todorov et al., 2005).

Another possibility is that context only moderates explicitly stated preferences. Trait judgments from faces occur spontaneously and quickly (Klapper et al., 2016; Willis & Todorov, 2006) and voters may be unaware of their influence. This suggests that the effect of trait impressions on voting may not be susceptible to voters' context-specific leader preferences. We do not think that this explanation for the current null results is likely though, as there is ample evidence showing that the effect of facial appearance on decisions varies across different contexts.

For example, preferences for dominant-looking partners are stronger when intergroup conflict is made salient (Hehman, Leitner, Deegan, & Gaertner, 2015). In the political domain, cultural differences (F. F. Chen, Jing, & Lee, 2012; Rule et al., 2010) and the political knowledge of voters (Berggren, Jordahl, & Poutvaara, 2017; Lenz & Lawson, 2011) have been shown to moderate how much voters rely on the facial appearance of candidates when making voting decisions. Previously mentioned work by Little and colleagues (2014; 2007, 2012) also demonstrates that leaders with different facial appearance are favored when an election is framed as taking place during a time of war or peace. In sum, these findings suggest that contextual factors can moderate the effect of trait perceptions on decision-making. More studies are needed to explore under what conditions election context moderates trait preferences in politicians.

Going beyond the political domain, there is ample evidence showing that trait impressions impact decision-making in various domains such as criminal sentencing (J. P. Wilson & Rule, 2015), personnel selection (Gomulya, Wong, Ormiston, & Boeker, 2017), and consumer behavior (Jaeger, Slegers, Evans, Stel, & van Beest, 2019). However, there is often conflicting evidence regarding which trait impressions people rely on. For example, Ert and colleagues (2016) analyzed Airbnb prices in Stockholm and found that people favor trustworthy-looking, but not attractive-looking hosts. The opposite pattern was found in a larger set of listing in New York City (Jaeger, Slegers, Evans, et al., 2019). Duarte and colleagues (2012) found that trustworthy-looking, but not attractive-looking loan applicants were more likely to be funded, whereas Ravina and colleagues (2008) did find an effect of attractiveness. More research is needed to understand which trait perceptions from faces predict different real-world outcomes. Future studies should consider a wide range of traits and test for potential moderating factors.

It is also plausible that different traits are preferred at different stages of the decision-making process. For example, Re and Rule (2017) found that faces of mafia members were perceived to be more powerful but less socially skilled than faces of lawyers, suggesting that different traits are valued in these two groups. However, this pattern reversed when analyzing rank attainment within groups: Perceived social skills were correlated with the rank of mafia members, while perceived power was correlated with the rank of lawyers. Thus, distinct traits were related to selection into a group and rank attainment within the group. In the context of political elections, it may be the case that certain traits are required to become a politician, to be nominated as a candidate, or to survive a preliminary round, whereas other traits are related to electoral success. It is unlikely though, that this feature can explain diverging results between the current study and previous investigations. Similar to previous studies showing that competent-looking politicians are more successful (Antonakis & Dalgas, 2009; Ballew & Todorov, 2007; Todorov et al., 2005), we analyzed results from the second round of run-off elections, but found no effect of perceived competence.

One shortcoming of the current set of studies was the limited number of constituencies. Our sample size was constrained by the number of constituencies that met our predefined inclusion criteria. For example, we decided to discard all elections for which we were not able to find a suitable photo for both candidates and all elections that featured at least one female candidate to avoid introducing gender stereotypes as a confound (Chiao et al., 2008). Our sample of 75 constituencies was larger than that of most studies which previously examined (and found) effects of facial appearance on election outcomes (e.g., Castelli et al., 2009; Chen, Jing, & Lee, 2012, 2014; Chen et al., 2016; Rule et al., 2010). Nonetheless, it might have been insufficient to detect regional differences in the effect of perceived trustworthiness. Future studies should consider a wider set of constituencies to ensure that the failure

to find support for our central hypothesis was not due to insufficient power.

### **Conclusion**

Even though we did not find any evidence that corruption moderates the success of trustworthy-looking politicians, our results do support the general notion that the facial appearance of political candidates is associated with voting behavior. In the context of the 2016 Italian local elections, attractive-looking candidates had a non-trivial advantage over their less attractive-looking opponents. Specifically, candidates who scored one standard deviation higher on perceived attractiveness received 2.98 percentage points more votes and were 1.91 times more likely to win. Thus, our results suggest that people rely on trait impressions from faces even when making such consequential decisions as whom to elect as their political leader.





# Chapter 3

The effects of facial attractiveness and trustworthiness in online peer-to-peer markets

Based on:

Jaeger, B., Slegers, W. W. A., Evans, A. M., Stel, M., & van Beest, I. (2019). The effects of facial attractiveness and trustworthiness in online peer-to-peer markets. *Journal of Economic Psychology*, 75, 102125.

All data, preregistration documents, and analysis scripts are available at the Open Science Framework (<https://osf.io/3enh8/>).

### **Abstract**

Online peer-to-peer markets, such as Airbnb, often include profile photos of sellers to reduce anonymity. Ert and colleagues (2016) found that more trustworthy-looking, but not more attractive-looking, Airbnb hosts from Stockholm charge higher prices for similar apartments. This suggests that people are willing to pay more for a night in an apartment if the host looks trustworthy. Here, we present a pre-registered replication testing how photo-based impressions of hosts' attractiveness and trustworthiness influence rental prices. We extend previous investigations by (a) controlling for additional features related to price (e.g., the apartment's location value), (b) testing for an influence of other host features, such as race and facial expression, and (c) analyzing a substantially larger sample of apartments. An analysis of 1,020 listings in New York City showed that more attractive-looking, but not more trustworthy-looking, hosts charge higher prices for their apartments. Compared to White hosts, Black (but not Asian) hosts charge lower prices for their apartments. Hosts who smile more intensely in their profile photo charge higher prices. Our results support the general conclusion that people rely on profile photos in online markets, though we find that attractiveness is more important than trustworthiness.

In recent years, online peer-to-peer markets such as Airbnb, eBay, and Uber have become increasingly popular. While these markets offer new opportunities for the exchange of goods and services, they also present a unique challenge. Unlike sellers in traditional markets, sellers on peer-to-peer platforms are not registered business, but private citizens. Sellers' reputations are relatively uncertain and purchases can be perceived as more risky. As a consequence, establishing trust between sellers and customers is a key challenge for peer-to-peer platforms (Einav, Farronato, & Levin, 2016). Building trust is particularly important for markets where the advertised service involves direct contact between consumers and sellers. On these platforms, sellers can provide services such as accommodation (e.g., Airbnb) or transport (e.g., Uber). However, people might be reluctant to enter the home or car of a complete stranger.

In order to facilitate trust between sellers and consumers, platforms include a variety of information about sellers. Next to review scores, profile photos are a common feature. Photos of sellers are meant to reduce anonymity, as well as facilitate identification offline (Guttentag, 2013). Critically, Ert and colleagues (2016) demonstrated that consumers on Airbnb use profile photos for more than identification—more trustworthy-looking hosts charge higher prices for similar apartments (i.e., when keeping other factors, such as review score and number of bedrooms constant). This suggests that consumers are willing to spend more on a night in an apartment when they perceive the host to be trustworthy. In other words, consumers seem to rely on first impressions based on a seller's profile photo when deciding which apartment to book.

Here, we aim to replicate and extend Ert and colleagues' (2016; Study 1) findings by controlling for additional features related to price of apartments, by testing for an influence of other photo-based

impressions of hosts, such as race and facial expression, and by analyzing a substantially larger sample of apartments from a different city.

### **Spontaneous trait inferences from faces**

People spontaneously infer personality characteristics of individuals solely based on their facial appearance (Todorov, Olivola, et al., 2015). Specifically, faces are evaluated on three dimensions: trustworthiness, dominance, and attractiveness (Sutherland, Liu, et al., 2017; Sutherland et al., 2013). In line with other models of person and group perception (Abele & Wojciszke, 2007; Fiske, Cuddy, Glick, & Xu, 2002; see also Sutherland, Oldmeadow, & Young, 2016), the first two dimensions on which faces are evaluated reflect evaluations of a target's intentions and abilities and are best captured by judgments of trustworthiness and dominance (Oosterhof & Todorov, 2008). The third dimension reflects an evaluation of the targets attractiveness (Sutherland, Liu, et al., 2017; Sutherland et al., 2013). People demonstrate some agreement in their face judgments (Hehman et al., 2017). However, while people might share stereotypes about what, for example, a trustworthy person looks like, their impressions have limited accuracy at best (Olivola & Todorov, 2010b; Rule et al., 2013; Todorov & Porter, 2014b).

### **The influence of first impressions in peer-to-peer markets**

The widespread use of profile photos in online peer-to-peer markets and the spontaneous nature of face judgments raises the question whether photo-based impressions of sellers influence people's decision-making in these markets. Given the important role of trust in peer-to-peer markets, one would expect sellers' perceived trustworthiness to play a central role (Guttentag, 2013; X. Ma, Hancock, Mingjie, & Naaman, 2017). However, perceived attractiveness has also been shown to influence decision-making in situations that are not

directly related to mate search. For example, attractive people receive more favorable treatment regarding personnel selection, career advancement, and wage distribution (Maestriperi et al., 2017). Therefore, both perceived attractiveness and trustworthiness of sellers might inform consumers' decisions in peer-to-peer markets.

Several studies have compared the effects of trustworthiness and attractiveness in peer-to-peer markets: Analyzing data from the crowdsourcing platform prosper.com, Duarte and colleagues (2012) found that more trustworthy-looking individuals are more likely to receive funding and receive more favorable interest rates; applicants' facial attractiveness was unrelated to their success. On the other hand, a similar study showed an advantage for attractive and creditworthy-looking borrowers (Ravina, 2008).<sup>9</sup>

In a similar vein, Ert and colleagues (2016; Study 1) investigated potential relationships between the perceived trustworthiness and attractiveness of Airbnb hosts and their apartment rental prices. If consumers favor trustworthy-looking or attractive-looking hosts, then those hosts should on average be able to rent out their apartments at higher prices. Thus, a preference for hosts with a certain facial appearance can be quantified by predicting the price of listings with the facial appearance of hosts and other characteristics that might be valued by consumers and therefore influence the price of a listing (cf. Rosen, 1974). A price analysis of 175 listings in Stockholm showed that—controlling for a variety of other features such as review score and whether or not the apartment is shared with the host—more trustworthy-looking hosts charge higher prices for similar apartments. No effect of attractiveness on apartment prices was found. This suggests

---

<sup>9</sup> Ravina (2008) found no effect of perceived trustworthiness. However, trustworthiness and creditworthiness were highly correlated. In the context of lending decisions, it is thus not surprising that trustworthiness did not predict the outcomes of lending decisions when perceived creditworthiness was accounted for.

that consumers favor trustworthy-looking hosts and are willing to pay higher prices to stay with them. Ert and colleagues (2016) also found a negative interaction between hosts' perceived trustworthiness and attractiveness. The more attractive-looking the host, the smaller the positive effect of perceived trustworthiness. Moreover, in a follow-up experiment, Ert and colleagues (2016) manipulated the perceived trustworthiness of hosts and the variance in review scores. In this context, both perceived trustworthiness and attractiveness predicted participant's apartment preferences.

In sum, findings from previous studies show that people rely on trait inferences from faces when making decision in peer-to-peer markets, even when they have access to other relevant information such as credit history or review scores. However, the current evidence on whether people favor attractive-looking or trustworthy-looking sellers is mixed.

### **Study 3.1**

We present a pre-registered replication study that builds on the findings by Ert and colleagues (2016; Study 1). Our goal is to provide a more comprehensive analysis of the influence of facial cues on consumer decisions on Airbnb. Specifically, our study design contains three notable improvements. First, we control for additional features that have been shown to influence the price of Airbnb listings: the attractiveness of the apartment's location and whether the host is a so-called superhost (Edelman & Luca, 2014; Gibbs, Guttentag, Gretzel, Morton, & Goodwill, 2018). Airbnb uses the superhost designation to highlight hosts who pass certain quality checks such as a high response rate and a low cancellation rate.

Second, we explore the influence of additional facial features (i.e., perceptions of the hosts' race, age, and smile intensity) on apartment prices. While the influence of some of these features has not been

explored yet, they also represent potential confounds for the effects of attractiveness or trustworthiness. For example, smiling is positively related to perceived trustworthiness and attractiveness (Sutherland, Young, & Rhodes, 2017). Controlling for these additional features provides a more robust test of the effects of facial attractiveness and trustworthiness on consumers' decisions. Third, we analyze a substantially larger sample of listings in New York City. Simonsohn (2015) suggested that, as a rule of thumb, replication studies should aim for a sample size that is at least 2.5 times larger than the original study. Here, we collect a sample of 1,020 listings, which is 5.8 times larger than the sample of the original study.

In addition to the analyses mentioned above, two exploratory analyses are presented. We investigate whether any effect of attractiveness is due to a beauty premium (i.e., more attractive hosts charging higher prices than hosts of average attractiveness), an ugliness penalty (i.e., less attractive charging less than hosts of average attractiveness), or both. This distinction is rarely tested in the literature and any effect of attractiveness is usually referred to as a beauty premium. Facial attractiveness is strongly correlated with perceptions of health (Jaeger et al., 2018; Pazda, Thorstenson, Elliot, & Perrett, 2016; Rhodes, 2006). Since people should be particularly motivated to avoid unhealthy (and therefore unattractive) individuals (Schaller & Duncan, 2007; Zebrowitz et al., 2003; Zebrowitz & Rhodes, 2004), this account would predict an ugliness penalty, but not necessarily a beauty premium.

Attractiveness biases might also be due to stereotypes linking attractiveness to more positive personality traits (Dion, Berscheid, & Walster, 1972; Eagly, Ashmore, Makhijani, & Longo, 1991). This account would predict a beauty premium, but not necessarily an ugliness penalty. Finally, we test for an effect of host race on apartment prices. This provides a replication of previous studies reporting that, compared to White host, Black and Asian hosts charge lower prices for similar

apartments (Edelman & Luca, 2014; Kakar, Franco, Voelz, & Wu, 2016; Wang, Xi, & Gilheany, 2015).

## Methods

This study was pre-registered and all data and analysis scripts are available at the Open Science Framework (<https://osf.io/3enh8/>). We report how our sample size was determined, all data exclusions, and all measures in the study.

### Airbnb data

We downloaded the New York City data set from the Inside Airbnb website (<http://insideairbnb.com>). This website features information on all Airbnb listings available in a specific city on a specific day. We selected New York City because it is one of the largest Airbnb markets worldwide. The dataset contains 40,227 Airbnb listings that were available on 3 December 2015.

Next, we applied our pre-registered exclusion criteria. Our analysis focused on apartments (as opposed to, for example, guesthouses or bed and breakfasts), as they represent the majority of advertised listings (86.22%). Apartments in which the rented room was shared with the host were also relatively rare (3.43%) and therefore excluded. We only selected apartments that were available for at least 30 days in the previous year and that received at least five reviews. The host of the apartment had to have a verified identity, a profile photo available, and only one listing for rent. We extracted the zip code of each listings and recorded the median rent for an apartment in that neighborhood.<sup>10</sup> This served as our measure of location value for each listing. Listings from zip codes with no available rent data were excluded (5,809 listings remaining).

---

<sup>10</sup> Rental data was accessed via [www.trulia.com](http://www.trulia.com). Values indicate the median rental price for an apartment in a given zip code in January 2017.



For the remaining listings, we downloaded the profile photos of hosts and selected the ones with only one depicted person in which the face of the host was clearly visible (2,359 listings remaining). We also downloaded the first photo of each listing which showed the apartment and selected the ones that give an impression of the inside living space (as opposed to, for example, photos of the New York City skyline; 2,110 listings remaining). Due to resource constraints, we randomly sampled 1,020 listings from the pool of remaining listings. Our analyses are based on this final sample of listings. For each listing, we recorded whether the entire apartment is rented out or shared with the host, whether the host is a superhost, the gender of the host<sup>11</sup>, number of bedrooms, median local rent, number of reviews, review score<sup>12</sup>, and the price per night.

### Photo ratings

We recruited 1,364 U. S. American workers from Amazon Mechanical Turk to rate 60 photos in exchange for 50 cents. We only recruited workers with approval rates above 90%. Data from ten participants who reported only poor or basic proficiency in English and data from 13 participants who always indicated the same rating was discarded, leaving a final sample of 1,336 participants ( $M_{\text{age}} = 36.22$ ,  $SD_{\text{age}} = 11.60$ ; 49.4% female). Participants were randomly allocated to one of three conditions, which determined what they rated the photos on: the trustworthiness of hosts ( $n = 446$ ), the attractiveness of hosts ( $n = 443$ ), or the attractiveness of apartments ( $n = 447$ ). Each participant rated the photos on only one trait in order to avoid consistency effects in ratings (cf. Penton-Voak, Pound, Little, & Perrett, 2006). On average, each photo was rated by 26 participants ( $Min = 23$ ,  $Max = 30$ ), which should be

---

<sup>11</sup> The first two authors independently coded the gender of all hosts by visually inspecting the profile photos. Agreement was at 100%.

<sup>12</sup> Note that review scores are displayed to users on a scale from 1 to 5 stars (rounded in increments of 0.5 stars). The review score variable, which reflects the listing's average review, ranges from 20 (1 star) to 100 (5 stars).

sufficient to obtain stable average ratings (Hehman et al., 2018). Participants were unaware that the photos were taken from Airbnb.

In the two host photo conditions, participants saw a random subset of 60 profile photos and were asked to rate the depicted person's trustworthiness or attractiveness on an 11-point scale ranging from *not at all* [trait] (0) to *extremely* [trait] (11). Each participant rated the photos on only one trait dimension. In the apartment condition, participants saw a random subset of 60 apartment photos and were asked to rate the attractiveness of the apartment on a similar scale. Following the procedure of Ert and colleagues (2016), we calculated the median ratings as our indicators of perceived trustworthiness and attractiveness of the hosts and perceived attractiveness of the apartments.

### **Photo classification**

Next, we used the Face++ application ([www.faceplusplus.com](http://www.faceplusplus.com)) to classify the hosts' race, age, and smile intensity. Face++ is a commercial algorithm that has been used in previous research to extract various indicators from large numbers of face images (Edelman, Luca, & Svirsky, 2017; Kosinski, 2017). For example, Edelman and colleagues (2017) used Face++ to classify the race of Airbnb guests. Face++ provides three race categorizations: White, Black, or Asian.<sup>13</sup> Face++ also provides a continuous age estimate, and a smile intensity score that ranges from 0 to 100. Past studies have found high accuracy levels for the classification of race and age (An & Weber, 2016; Jaeger, Slegers, & Evans, 2019; Rhue & Clark, 2016).

---

<sup>13</sup> Naturally, the algorithm's classification is only based on superficial perceptual cues that can be extracted from a photograph, such as face shape and skin color. We do not claim that the algorithm's broad classification provides an accurate reflection of an individual's ethnic background. However, this is not a limitation in the current context as we were not interested in the influence of a hosts' actual race or ethnicity, but rather in their race category as perceived by consumers on the basis of a profile photo.

## Results

Price, median local rent, and number of reviews were  $\log_{10}$ -transformed due to their skewed distributions and all continuous variables were z-standardized. We entered all variables into OLS regression models with price as the outcome variable.

### Descriptive statistics

The price per night of listings ranged from \$25 to \$1,500 with a median price of \$128.50 ( $M = \$150.62$ ,  $SD = \$104.10$ ). The perceived trustworthiness of hosts ranged from 2 to 9 on our 11-point scale ( $M = 6.00$ ,  $SD = 1.06$ ) and the same spread was observed for the perceived attractiveness of hosts ( $M = 5.58$ ,  $SD = 1.30$ ). We found a small correlation between perceived trustworthiness and perceived attractiveness,  $r(1,018) = .22$ ,  $p < .001$ .

Face++ was unable to provide classifications for three hosts (0.29%). Of the remaining 1,017 hosts, 73.84% were classified as White, 12.49% as Black, and 13.67% as Asian. There was a significant effect of race on perceived attractiveness,  $F(2, 1,014) = 6.87$ ,  $p = .001$ , but not on perceived trustworthiness,  $F(2, 1,014) = 1.84$ ,  $p = .16$ . Participants rated Black hosts ( $M = 5.21$ ,  $SD = 1.11$ ) as less attractive than White hosts ( $M = 5.66$ ,  $SD = 1.28$ ),  $t(186.9) = 4.11$ ,  $p < .001$ ,  $d = 0.39$ , and marginally less attractive than Asian hosts ( $M = 5.51$ ,  $SD = 1.47$ ),  $t(255.3) = 1.88$ ,  $p = .062$ ,  $d = 0.50$ . We did not find a significant difference in attractiveness between White and Asian hosts,  $t(186.9) = 1.13$ ,  $p = .26$ ,  $d = 0.10$ .

Age was negatively correlated with attractiveness,  $r(1,015) = .25$ ,  $p < .001$ , but there was no significant correlation with trustworthiness,  $r(1,015) = -.0008$ ,  $p = .98$ . Smile intensity was positively correlated with attractiveness,  $r(1,015) = .13$ ,  $p < .001$ , and with trustworthiness,  $r(1,015) = .47$ ,  $p < .001$ . Descriptive statistics for all predictors can be found in Table 3.1 (for continuous variables) and Table 3.2 (for

## Chapter 3

categorical variables). For ease of interpretation, we report descriptive statistics if unstandardized variables.

Table 3.1  
*Descriptive statistics for all continuous variables.*

Variable	M	SD	Min	Max	Median	Skew
Trustworthiness	6.00	1.06	2.00	9.00	6.00	-0.486
Attractiveness	5.58	1.30	2.00	9.00	6.00	-0.105
Price	150.62	104.10	25.00	1,500.00	128.50	4.322
Apartment photo	5.84	1.41	2.00	10.00	6.00	-0.228
Bedrooms	1.08	0.61	0.00	4.00	1.00	1.502
Median local rent	4,517.69	3,948.06	650.00	45,000.00	3,399.00	4.771
# of reviews	31.96	32.40	5.00	251.00	19.00	2.233
Review score	93.29	5.35	65.00	100.00	94.00	-1.318
Face++ smile	57.98	36.59	0.14	99.83	69.53	-0.311
Face++ age	43.08	10.21	12.00	72.00	44.00	-0.098

Table 3.2  
*Descriptive statistics for all categorical variables.*

Variable	Group	N	%
Entire apartment	No	387	37.94
	Yes	633	62.06
Host gender	Female	529	51.86
	Male	491	48.14
Superhost	No	863	84.61
	Yes	157	15.39
Face++ race	Black	127	12.45
	Asian	139	13.63
	White	751	73.63
	Undetected	3	0.29

### Confirmatory analyses

In accordance with the approach by Ert and colleagues (2016), we conducted multiple regression analyses to predict the price of listings with the hosts' perceived trustworthiness and attractiveness, while controlling for other features of the host and the listing (see Table 3.3). We did not find an effect of host trustworthiness,  $\beta = -0.0036$ ,  $SE = 0.0047$ ,  $t(1,010) = -0.76$ ,  $p = .45$ , 95% CI [-0.013, 0.0057] (Model 1). However, we did find a positive effect of host attractiveness,  $\beta = 0.011$ ,  $SE = 0.0048$ ,  $t(1,010) = 2.34$ ,  $p = .020$ , 95% CI [0.0018, 0.021] (Model 2). These effects did not change when host attractiveness and trustworthiness were entered simultaneously into a model (Model 3). There was no significant interaction between host attractiveness and trustworthiness,  $\beta = 0.0075$ ,  $SE = 0.0042$ ,  $t(1,008) = 1.79$ ,  $p = .074$ , 95% CI [-0.00074, 0.016] (Model 4).

Thus, in contrast to Ert and colleagues (2016; Study 1), we did not find a main effect of perceived trustworthiness, or an interaction effect between trustworthiness and attractiveness. Instead, we found a main effect of attractiveness. Specifically, a one standard deviation increase in perceived attractiveness was associated with a 2.78% price increase. As a comparison, a one standard deviation increase in review score was associated with a 5.26% price increase and the presence of an additional bedroom (which can also be seen as a proxy for the apartment's size) was associated with a 15.66% price increase.

Table 3.3  
*The influence of facial trustworthiness and attractiveness on the price of Airbnb listings.*

	Model 1	Model 2	Model 3	Model 4
Trustworthiness	-0.004		-0.005	-0.004
Attractiveness		0.011*	0.012*	0.013**
Trustw.*Attract.				0.008
Apartment rating	0.054***	0.053***	0.053***	0.053***
Bedrooms	0.063***	0.063***	0.063***	0.063***
Entire apartment	0.246***	0.245***	0.245***	0.246***
Male host	-0.002	0.008	0.005	0.005
Superhost	0.013	0.014	0.015	0.015
Median local rent	0.077***	0.076***	0.076***	0.076***
# of reviews	0.001	0.002	0.002	0.002
Review score	0.022***	0.022***	0.022***	0.022***
adj. R <sup>2</sup>	.636	.637	.637	.638

\*\*\*  $p < .001$ . \*\*  $p < .01$ . \*  $p < .05$ .



Next, we tested for potential interaction effects between the facial appearance of hosts and other characteristics of the hosts or their apartments. Physical attractiveness is valued more in women than men (Buss, 1989; Feingold, 1990) and financial benefits for more attractive individuals have been observed more consistently when women rather than men were the targets (Maestripieri et al., 2017). However, we found no interactions between the gender of the host and their perceived attractiveness,  $\beta = 0.0057$ ,  $SE = 0.0096$ ,  $t(1,008) = 0.59$ ,  $p = .56$ , 95% CI [-0.013, 0.025], or their perceived trustworthiness,  $\beta = -0.018$ ,  $SE = 0.0095$ ,  $t(1,008) = -1.89$ ,  $p = .058$ , 95% CI [-0.037, 0.00065].

We also tested for potential interaction effects between perceived attractiveness or trustworthiness and whether the entire apartment is rented out rather than shared with the host. It is likely that consumers who share the apartment have more direct contact with the host. They might therefore be more concerned with selecting a desirable host. However, we found no interaction between whether the entire apartment was offered (vs. shared with the host) and the host's perceived attractiveness,  $\beta = -0.0016$ ,  $SE = 0.0096$ ,  $t(1,008) = -0.17$ ,  $p = .86$ , 95% CI [-0.02, 0.017], or perceived trustworthiness,  $\beta = -0.0077$ ,  $SE = 0.0091$ ,  $t(1,008) = -0.85$ ,  $p = .40$ , 95% CI [-0.026, 0.01].<sup>14</sup>

### Exploratory analyses

**Beauty premium vs. ugliness penalty.** To test for independent effects of low and high attractiveness, we grouped hosts into three attractiveness categories: one standard deviation below average or lower ( $n = 204$ , 20%), one standard deviation above average or higher ( $n = 253$ , 24.80%), or in between ( $n = 563$ , 55.20%). Regressing price on

---

<sup>14</sup> We repeated all analyses reported here with mean trustworthiness and attractiveness ratings, as opposed to median ratings, but no differences in results were found. In a similar vein, excluding six listings with prices that were three or more standard deviations above or below the mean log-transformed price led to qualitatively equivalent results.

attractiveness category (with average attractiveness as the reference group) revealed evidence for an ugliness penalty, but not a beauty premium (Table 3.4, Model 5): Low attractiveness was negatively related to price,  $\beta = -0.031$ ,  $SE = 0.012$ ,  $t(1,008) = -2.60$ ,  $p = .009$ , 95% CI [-0.054, -0.0075], whereas we did not find evidence that high attractiveness was positively related to price,  $\beta = 0.0028$ ,  $SE = 0.011$ ,  $t(1,008) = 0.25$ ,  $p = .80$ , 95% CI [-0.019, 0.025]. Specifically, relatively unattractive hosts charged 6.82% less for their listings.

**The influence of race, age, and smile intensity.** We also extended our analysis by including additional characteristics acquired through the Face++ algorithm, which was used to classify a host's race, age, and smile intensity based on their profile photo. Including race, age, and smile intensity in our regression model showed that, compared to White hosts, Black hosts charged significantly lower prices for their listings,  $\beta = -0.046$ ,  $SE = 0.014$ ,  $t(1,002) = -3.32$ ,  $p < .001$ , 95% CI [-0.074, -0.019] (Table 3.4, Model 6). Specifically, Black hosts charged 10.09% lower prices for similar apartments. We found no price difference between White and Asian hosts,  $\beta = 0.0025$ ,  $SE = 0.013$ ,  $t(1,002) = 0.19$ ,  $p = .85$ , 95% CI [-0.024, 0.029]. Furthermore, estimated age of the host was not associated with the price of their apartment,  $\beta = 0.0051$ ,  $SE = 0.0048$ ,  $t(1,002) = 1.07$ ,  $p = .28$ , 95% CI [-0.0042, 0.014].

The smile intensity of the host was positively associated with rental price,  $\beta = 0.015$ ,  $SE = 0.0051$ ,  $t(1,002) = 3.00$ ,  $p = .003$ , 95% CI [0.0053, 0.025]. A one standard deviation increase in smile intensity was related to a 3.61% price increase. Crucially, we still found a positive effect of perceived attractiveness when controlling for these additional variables  $\beta = 0.011$ ,  $SE = 0.0050$ ,  $t(1,002) = 2.34$ ,  $p = .019$ , 95% CI [0.0019, 0.021], showing that the positive effect of perceived attractiveness is not due to the host's race, age, or smile intensity. We also found a negative effect of perceived trustworthiness  $\beta = -0.012$ ,  $SE = 0.0053$ ,  $t(1,002) = 2.32$ ,  $p = .021$ , 95% CI [-0.023, -0.0019].

Table 3.4

*The influence of facial trustworthiness and attractiveness on the price of Airbnb listings when controlling for additional characteristics of the hosts.*

	Model 5	Model 6
Trustworthiness	-0.006	-0.012 *
Attractiveness		0.012 *
Low Attractiveness	-0.031 **	
High Attractiveness	0.003	
Black host		-0.046 ***
Asian host		0.002
Host age		0.005
Smile intensity		0.015 **
Apartment rating	0.053 ***	0.054 ***
Bedrooms	0.063 ***	0.063 ***
Entire apartment	0.246 ***	0.247 ***
Male host	0.001	0.006
Superhost	0.016	0.013
Median local rent	0.076 ***	0.072 ***
# of reviews (log)	0.001	0.002
Review score	0.022 ***	0.021 ***
adj. R <sup>2</sup>	.638	.643

\*\*\*  $p < .001$ . \*\*  $p < .01$ . \*  $p < .05$ . †  $p < .10$ .

### General discussion

We examined the relationship between perceived facial attractiveness and trustworthiness of Airbnb hosts and the price they are charging for their apartments. While some studies have shown that consumer decisions in online peer-to-peer markets are influenced by the

attractiveness and trustworthiness of sellers, evidence on which trait is favored in sellers is mixed (Duarte et al., 2012; Ert et al., 2016; Ravina, 2008). Our analysis of 1,020 Airbnb listings in New York City revealed that more attractive-looking hosts charge 2.78% higher prices for similar apartments. We did not find that more trustworthy-looking hosts charge different prices. This suggests that consumers are willing to spend more on apartments that are offered by more attractive host.

Our results do not replicate findings by Ert and colleagues (2016; Study 1) who reported a positive effect of perceived trustworthiness, but not attractiveness, in a sample of 175 listings. It should be noted that our study differed from this previous investigation in a few notable ways. First, we controlled for additional factors related to the apartment (the attractiveness of the apartment's location) and the host (race, age, smile intensity, and whether they are a superhost) that could confound the relationship between photo-based impressions and the price of listings.

Second, we analyzed a substantially larger sample ( $n = 1,020$ ). Both should result in a more precise estimate of the influence of facial attractiveness and trustworthiness on apartment prices. Third, we analyzed Airbnb listings from New York City rather than Stockholm in order to be able to collect a larger sample. It is possible that trait preferences for Airbnb hosts vary across different countries or cities. In fact, the association between other host characteristics (e.g., the superhost designation) and apartment prices has been shown to differ across different cities (Gibbs et al., 2018). Although, we have no theory-based explanation regarding the relative importance of perceived attractiveness and trustworthiness in Airbnb hosts in New York City and Stockholm, future studies could explore if reliable differences in trait preferences exist.

Our findings converge with previous studies showing that trait inferences from faces can influence a variety of decisions (Maestriperi et al., 2017; Olivola, Funk, et al., 2014). Forming trait impressions from

faces is a fast and intuitive process (Klapper et al., 2016; Ritchie et al., 2017; Willis & Todorov, 2006). Given the prominent role of profile photos in online peer-to-peer market, it may thus not seem surprising that the facial appearance of hosts influences people's decision whom to stay with. Importantly, the effect of facial attractiveness was observed when controlling for other photo-based cues such as the host's gender, age, race, and facial expression. In sum, we conclude that impressions of attractiveness guide consumer decisions even when a myriad of other cues are available.

Examining the influence of these other cues, we found that Black hosts charge on average 10.09% less for their apartments. This in line with the price gap of approximately 12% reported by Edelman and Luca (2014). However, in contrast to previous studies that analyzed Airbnb listing in San Francisco, Oakland, and Berkeley (Kakar et al., 2016; Wang et al., 2015), we did not find that Asian hosts charge less for their apartments. We also found that hosts with more intense smiles charge 3.61% higher rents for their apartments. Fagerstrøm and colleagues (2017) showed that participants were more motivated to explore an Airbnb listing's web page when the profile photo showed a smiling rather than a neutral host. Our results demonstrate that this preference for smiling hosts can also be observed in consumer's revealed preferences.

Why do consumers prefer to stay with attractive hosts? One explanation is that due to the critical importance of engaging with healthy partners, preferences for attractive individuals, consciously or unconsciously, spill over to situation that are not directly related to mate search (Maestripieri et al., 2017). Unattractive hosts might elicit avoidance motivations, which then spill over to consumer's apartment choices. In fact, similar effects of facial attractiveness on seemingly unrelated preferences have been observed for interest in scientific work (Gheorghiu, Callan, & Skylark, 2017). This account is also supported by

our current finding that the effect of attractiveness on apartment prices is driven by unattractive hosts charging lower prices, rather than by attractive hosts charging higher prices. Attractive individuals are seen as more healthy (Rhodes et al., 2007). Given the importance of avoiding unhealthy individuals (Schaller & Duncan, 2007; Zebrowitz et al., 2003; Zebrowitz & Rhodes, 2004), we would therefore expect the negative effect of low attractiveness to be stronger than the positive effect of high attractiveness (cf. Jaeger, Wagemans, et al., 2018; Pazda et al., 2016).

It is also possible that people consciously select attractive hosts because they believe that they will enjoy their stay with them more. For example, attractive people are believed to possess more positive personality traits (Dion et al., 1972; Eagly et al., 1991). While perceived trustworthiness did not predict apartment prices, people might value other traits such as sociability and rely on a host's attractiveness to infer their sociability. Do people actually have better stays with attractive hosts? We can test this by probing for a relationship between the attractiveness of hosts and their review score, which should reflect people's satisfaction with their stay. We do not find any evidence that people assign higher review scores to more attractive hosts,  $r(1,018) = -.019, p = .54$ .

Relatedly, if people consciously select attractive hosts out of sexual interest, we would expect a larger attractiveness effect when the apartment is shared and there is actual contact between guest and host. This prediction was not confirmed by our results either. In sum, the hypotheses that consumers consciously prefer attractive hosts due to sexual interest or that they derive pleasure from interacting with an attractive host is not supported by the current data. However, it should be noted that our study was not designed to test these different accounts directly and the results reported here should only be taken as preliminary evidence. Future studies need to address the exact reasons underlying consumers' reliance on attractiveness when deciding whom

to stay with. For example, if mating motives really play a role, we would expect larger effects of attractiveness in opposite-sex, rather than same-sex interactions. More generally, future studies on the influence of attractiveness would benefit from testing for different effects of low and high (vs. average) attractiveness. Currently, this distinction is rarely made in the literature and any effect of attractiveness is referred to as a beauty premium (e.g., Berggren, Jordahl, & Poutvaara, 2010; Gonzalez & Loureiro, 2014; Hamermesh & Biddle, 1994).

### **Limitations**

A potential limitation of the current study is the use of price as a proxy for consumer preferences. In line with previous studies (e.g., Edelman & Luca, 2014; Ert et al., 2016), we reasoned that, if consumers favor a particularly attractive-looking or trustworthy-looking host, then those hosts should be able to charge higher prices for their apartments (Malpezzi, 2008; Rosen, 1974). One requirement for such an analysis is that other factors, which are valued by consumers and thus drive the price, are controlled for. To this end, we included various characteristics of the apartment and the host in our models. This also allowed us to test the validity of our methodological approach. If price is determined by the presence or absence of features that are valued by consumers, then we would predict to find significant effects for features that should be strongly valued by people who are looking to rent an apartment on Airbnb. Indeed, the price of apartments was related to their size, review score, and location value. In other words, our results confirm the intuition that consumers would be willing to pay more for a larger apartment, for an apartment that received better reviews, or for an apartment that is located in a better neighborhood (see also Gibbs et al., 2018).

We also acknowledge that our work is correlation, which precludes us from making any causal claims. While experimental studies in the lab

provide more opportunities to disentangle the unique effects of different factors, analyzing real-world data such as the prices of Airbnb listings has the advantage of revealing actual behavior in an ecologically valid environment. Ultimately, we believe that evidence from both inside and outside the lab is needed to provide a convincing test of the influence of facial features on decision-making (Baumeister, Vohs, & Funder, 2007; Maner, 2016).

### **Practical implications**

Next to demonstrating the effect of facial attractiveness using real-world data, our findings have implications for the design of online peer-to-peer platforms. Edelman and Luca (2014)—who were the first to find evidence of a price gap between White and Black hosts on Airbnb, a finding which we successfully replicated—suggested that photos should be omitted from platforms in order to prevent racial discrimination. We also show that consumers not only discriminate on the basis of race, but also attractiveness. These findings might prompt some hosts to remove their profile photos to guard themselves against any appearance-based discrimination. However, we would be careful in advising Black or unattractive hosts to delete their photos unilaterally. People are less motivated to explore a listing's web page if no photo is displayed (and other listings include photos; Fagerstrøm et al., 2017) and they generally value photos in trust-based economic exchange (Eckel & Petrie, 2011; see also Heyes & List, 2016). Thus, removing one's profile photo might actually result in a similar price penalty. Future studies should investigate how the presence or absence of a profile photo influences preferences for sellers in peer-to-peer markets.

Similar to Edelman and Luca (2014), we advise platforms such as Airbnb to regulate which information about sellers is provided (and at what time). Photos enable personal identification, which can facilitate initial trust between sellers and consumers. However, this information



is not necessarily needed when consumers are browsing for apartments. Platforms could provide photos of sellers only at the moment a transaction has been made, or when users send initial inquiries to hosts about listings. This change would enable personal identification of sellers, but prevent consumers from engaging in appearance-based discrimination when selecting a rental location.

An alternative approach would be to increase the salience of objective rental information and decrease the salience of profile photos. Consumers pay more attention to information that takes up a lot of space (Wedel & Pieters, 2007). Many platforms currently display profile photos very prominently, which makes the seller's appearance a particularly salient feature. Instead of showing profile photos in large size at the top of a listing's web page, they could be displayed in reduced size on a separate page, such as the seller's personal profile. For example, Airbnb's most recent web design, which was implemented at the end of 2016, features a large photo of the listing, while the size of the host's photo (but not the prominence of its position) was reduced. In general, more data is needed to systematically test the influence of profile photos on consumer behavior. Future studies could test whether less salient photos actually reduces appearance-based discrimination and whether providing photos only after a transaction has been made affects.

## Conclusion

Our analysis of 1,020 apartments in New York City shows that more attractive-looking hosts charge higher prices for similar apartments, suggesting that consumers are willing to spend more on a night in an apartment if they are staying with a more attractive host. This effect was due to an ugliness penalty, rather than a beauty premium: Less attractive hosts charge lower prices whereas more attractive hosts do not charge higher prices. We did not replicate Ert and colleagues' (2016; Study 1) finding that the perceived trustworthiness of hosts influences apartment

## Chapter 3

prices. However, we did replicate previous findings showing that Black hosts charge lower prices for their apartments (Edelman & Luca, 2014). Taken together, our findings show that photo-based impressions guide consumer decisions in peer-to-peer markets.

# Chapter 4

Who judges a book by its cover? The prevalence, structure, and correlates of beliefs in physiognomy

Based on:

Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2019). *Who judges a book by its cover? The prevalence, structure, and correlates of beliefs in physiognomy*. Manuscript submitted for publication.

All data, analysis scripts, and preregistration documents are available at the Open Science Framework (<https://osf.io/s9nj8/>).

## Abstract

The question of whether personality can be inferred from faces is contentiously debated. We propose that, irrespective of the actual accuracy of trait inferences from faces, lay beliefs about the manifestation of personality traits in facial features (i.e., physiognomic beliefs) have important consequences for social cognition and behavior. In five studies ( $N = 3,861$ ), we examine the prevalence, structure, and correlates of physiognomic beliefs. We find that belief in physiognomy is common among students (Study 4.1) and in a large, representative sample of the Dutch population (Study 4.2). Physiognomic beliefs are relatively stable over time and associated with an intuitive thinking style (Study 4.3). However, the strength of physiognomic beliefs varies across different personality dimensions: sociability is believed to be more reflected in facial appearance than morality or competence (Studies 4.1-4.5). Crucially, individual differences in belief strength predict how people form and use first impressions. People with stronger physiognomic beliefs are more confident in their trait inferences (Study 4.4) and rely more on them when making decisions (Study 4.5). Yet, this increased confidence is not explained by superior accuracy of personality inferences, and the endorsement of physiognomic beliefs is associated with *overconfidence* (Study 4.4). Overall, there is widespread belief in physiognomy among laypeople, and individual differences in belief strength relate to various social-cognitive processes and behaviors.

The practice of physiognomy involves inferring psychological characteristics from facial (or other bodily) features. The idea dates back to Ancient Greece and enjoyed particular popularity in the 18<sup>th</sup> and 19<sup>th</sup> centuries (Aristotle, trans. 1936; Lavater, 1775; Woods, 2017). More rigorous scientific approaches at the beginning of the 20<sup>th</sup> century provided evidence against many of physiognomy's claims (e.g., Cleeton & Knight, 1924) and it is now widely regarded as pseudo-science (Todorov, 2017). However, research in the field of social perception has shown that faces play an important role in everyday impression formation: People spontaneously infer personality traits from facial appearance and these inferences guide many consequential decisions (Todorov, Olivola, et al., 2015). This raises questions about belief in physiognomy among laypeople.

People develop and rely on lay theories—systems of implicit or explicit beliefs—to navigate the social world (Baumeister & Monroe, 2014; Boyer & Petersen, 2018). For example, research on lay personality theory has shown that people hold beliefs about the basis (Haslam, Bastian, & Bissett, 2004), malleability (Chiu, Hong, & Dweck, 1997), structure (Stolier, Hehman, Keller, Walker, & Freeman, 2018), and expression (Mehl, Gosling, & Pennebaker, 2006) of personality traits. Critically, individual differences in these beliefs predict outcomes related to impression formation (Haslam et al., 2004), information search (Plaks, Stroessner, Dweck, & Sherman, 2001), and stereotyping (Levy, Stroessner, & Dweck, 1998). In other words, lay beliefs about different aspects of personality are widespread and shape various social-cognitive processes and behaviors.

Here, we examine a facet of lay personality theory that has received little attention thus far: the belief that personality is reflected in facial appearance (i.e., *physiognomic beliefs*). We propose that individual differences in physiognomic beliefs influence various aspects of impression formation, such as confidence in trait inferences and reliance

on trait inferences in social decision-making. We examine (a) the prevalence of physiognomic beliefs, (b) their structure (i.e., which characteristics people believe are most reflected in facial features), (c) the psychological correlates of physiognomic beliefs (e.g., the relationship between physiognomic beliefs and epistemic motivation), and (d) whether physiognomic beliefs are related to accurate social perception and greater reliance on facial appearance in social decision-making.

### **Physiognomy and social perception**

The core tenet of physiognomy holds that facial morphology (i.e., features of resting, non-expressive faces) is indicative of psychological characteristics and behavioral tendencies (Todorov, 2017). Early writings proposed that the size and orientation of facial features reflect their frequent use (e.g., a disagreeable person who frowns a lot will have lowered eyebrows); moreover, resemblances between humans and other animals were thought to point to shared psychological attributes (e.g., a person who looks like a lion is brave like a lion; Aristotle, trans. 1936). These speculations were not based on rigorous scientific study and many claimed links between specific facial features and personality traits were disconfirmed by empirical work at the beginning of the 20<sup>th</sup> century (Alley, 1988). For example, Cleeton and Knight (1924, p. 216) reported that the correlation between "variations in physical traits purported to reveal variations in character traits and [character] criteria was 0.000."

Even though early 20<sup>th</sup> century research found little support for physiognomy, interest in the topic has grown again in recent years. Research in the field of social perception has yielded new insights into the determinants of impression formation, showing that people spontaneously infer a variety of personality traits from resting, non-expressive faces (Todorov, Olivola, et al., 2015). In fact, people are

relatively confident in their own physiognomic judgments (Ames et al., 2010; Hassin & Trope, 2000) and rely on them when making a wide range of consequential decisions (Olivola, Funk, et al., 2014). For instance, voting decisions are influenced by the perceived competence of political candidates (Olivola & Todorov, 2010a) and criminal sentencing decisions are influenced by the perceived trustworthiness of defendants (J. P. Wilson & Rule, 2015). People even rely on facial appearance when they have access to superior, objective information (Jaeger, Evans, Stel, & van Beest, 2019a; Olivola et al., 2018). The pervasive influence of trait impressions has again raised questions about the diagnosticity of facial features for inferring personality (Bonnefon et al., 2017; Todorov, Olivola, et al., 2015; Zebrowitz et al., 2003)

Can people form accurate personality impressions based on facial appearance? To address this question, studies have tested whether personality impressions from faces (i.e., trait judgments based on facial photographs) reflect a target's actual personality (Berry, 1990; Penton-Voak et al., 2006). Overall, evidence in favor of accuracy in face-based personality judgments is weak and inconsistent. Some studies find a small "kernel of truth" in trait impressions (Bonnefon et al., 2013; De Neys et al., 2017; C. Lin et al., 2018; Penton-Voak et al., 2006; Satchell, Davis, Julle-Danière, Tupper, & Marshman, 2018; Slepian & Ames, 2015; Tognetti et al., 2013). However, others find no accuracy (Efferson & Vogt, 2013; Graham et al., 2017; Ling et al., 2019; Rule et al., 2013) or provide theoretical arguments against accurate trait impressions (McCullough & Reed, 2016; Todorov & Porter, 2014a). For example, trait impressions vary substantially across different perceivers (Helman et al., 2017), contexts (Brambilla et al., 2018), and even across different images of the same target (Todorov & Porter, 2014a), suggesting that they are not reliable indicators of personality. In short, the available evidence suggests that facial features are, at best, a weak indicator of personality.

## **Physiognomic beliefs**

Here, we propose that, irrespective of the actual accuracy of trait inferences from faces, people may hold lay beliefs about the manifestation of personality traits in facial features (i.e., physiognomic beliefs). Crucially, widespread beliefs in physiognomy in the general population may explain the pervasive effects of facial appearance on social cognition and behavior. In a similar vein, individual differences in physiognomic beliefs may predict why some people are overconfident in the accuracy of their trait impressions and persistently rely on them when making decisions.

Little is known about lay beliefs in physiognomy. Anecdotal evidence suggests that some people hold physiognomic beliefs (Hassin & Trope, 2000; Liggett, 1974). Moreover, a recent study by Suzuki and colleagues (2017) showed that physiognomic beliefs are related to (a) other lay beliefs, such as belief in a just world and belief in the biological determinism of personality and (b) more extreme trait judgments based on facial appearance.

The present research examines the role of physiognomic beliefs in social perception more broadly. First, we examine the prevalence of physiognomic beliefs in the general population. We assess belief in physiognomy in five samples (including a large, representative sample of the Dutch population) and explore who is more likely to endorse physiognomic beliefs. For example, we test whether physiognomic beliefs are correlated with various characteristics such as age, education level, or thinking style. Second, we investigate the heterogeneity of physiognomic beliefs across different personality dimensions. Specifically, we ask whether some personality traits are believed to be more reflected in faces than others. Third, and most importantly, we test whether individual differences in physiognomic beliefs predict how confident people are in the accuracy of their trait impressions and how much they rely on trait impressions in social decision-making.



## The current studies

To investigate the prevalence, structure, and correlates of physiognomic beliefs, we introduce a novel scale.<sup>15</sup> The scale consists of two parts with a total of 15 items. To ensure that participants envision a resting, non-expressive face, we prompt them to “imagine seeing the passport photo of a stranger”. The first part (3 items) assesses *general physiognomic beliefs* (e.g., “I can learn something about the person’s personality just from looking at his or her face”). The second part (12 items) assesses *specific physiognomic beliefs* by asking respondents to indicate how accurately they think different characteristics can be inferred from a person’s face. Our scale measures physiognomic beliefs for three fundamental dimensions underlying person perception (sociability, morality, and competence; Brambilla, Rusconi, Sacchi, & Cherubini, 2011; Leach, Ellemers, & Barreto, 2007), as well as three additional characteristics (age, gender, and attractiveness).<sup>16</sup> This allows us to test which characteristics are believed to be more reflected in faces.

We report the results of five studies ( $N = 3,861$ ; four preregistered). In Studies 4.1 and 4.2, we investigate the prevalence and structure of physiognomic beliefs in a student sample ( $n = 378$ ) and a representative sample of the Dutch population ( $n = 2,624$ ). We estimate how many

---

<sup>15</sup> Critically, our measure differs from the measure introduced by Suzuki and colleagues (2017), which assessed physiognomic beliefs by asking participants which characteristics (e.g., aggressiveness, cooperativeness) they think they can determine about a person from their face. It is unclear whether high scores on their scale reflect beliefs that traits can be inferred from stable, morphological features of faces—which is the claim of physiognomy (Aristotle, 1936; Lavater, 1775) and the subject of debate (e.g., Bonnefon et al., 2015; Todorov, Funk, et al., 2015)—or from other facial characteristics such as emotion expressions.

<sup>16</sup> Research on the accuracy of trait inferences from faces has predominantly focused on basic personality dimensions such as the Big Five traits (Penton-Voak et al., 2006). Since we were interested in people’s conceptual beliefs about personality, we focused on sociability, morality, and competence. These dimensions may not capture the structure of a person’s actual personality as well as other models, but they represent the dimensions that people spontaneously use to judge others’ personality (Brambilla et al., 2011).

people believe that personality is reflected in facial features and test which personality dimensions are believed to be more visible in faces. Study 4.3 ( $n = 229$ ) examines the relationship between physiognomic beliefs and individual differences in other lay beliefs (e.g., belief in the biological determinism of personality traits; Haslam et al., 2004) and epistemic motivation (e.g., faith in intuition; Epstein, Pacini, Denes-Raj, & Heier, 1996). We also examine the temporal stability of physiognomic beliefs.

Studies 4.4 and 4.5 investigate the relationships between physiognomic beliefs and social perception and decision-making. Study 4.4 ( $n = 406$ ) tests whether people who score higher on physiognomic beliefs are more confident in the accuracy of their trait impressions and whether increased confidence can be explained by the fact that their judgments are indeed more accurate. Finally, Study 4.5 ( $n = 224$ ) investigates whether people who endorse physiognomic beliefs rely more on trait inferences from faces when making social decisions.

All data, analysis scripts, and preregistration documents are available at the Open Science Framework (<https://osf.io/s9nj8/>). We report how our sample sizes were determined and all data exclusions and measures for each study.

### **Studies 4.1 and 4.2: Prevalence and structure**

In Studies 4.1 and 4.2, we estimated the prevalence of physiognomic beliefs and tested how belief strength varies for different personality traits. We predicted that facial appearance would be seen as more indicative of physically salient characteristics, such as gender, age, and attractiveness. People can detect gender and age from faces with high accuracy (Bruce & Young, 2012), and overall attractiveness is strongly influenced by facial appearance (M. Peters, Rhodes, & Simmons, 2007). We also examined differences in physiognomic beliefs across three fundamental dimensions in person perception: sociability,

morality, and competence. We predicted that people would hold stronger physiognomic beliefs for sociability (compared to morality and competence). Emotion perception in resting (i.e., emotionally neutral) faces play a central role in impression formation (Said et al., 2009). Critically, emotional expressiveness is a defining feature of sociability, making facial appearance particularly relevant for sociability judgments (Kring, Smith, & Neale, 1994; Riggio & Riggio, 2002).

We tested our predictions by administering the physiognomic belief scale in a sample of first-year psychology students (Study 4.1;  $n = 378$ ) and a representative sample of the Dutch population (Study 4.2;  $n = 2,624$ ).

## Methods

**Participants.** In Study 4.1, we recruited 378 first-year psychology students from a Dutch university ( $M_{age} = 20.61$ ,  $SD_{age} = 2.19$ ; 76.46% female, 23.28% male, 0.26% other). The majority of participants were Dutch (68.25%) or German (19.05%). Sample size was determined by how many students participated in the study within two weeks. A sensitivity analysis in G\*Power (Faul et al., 2007) showed that this sample size sample afforded us 80% power to detect a small difference ( $d = 0.14$ ) when comparing physiognomic beliefs for sociability, morality, and competence (with  $\alpha = 5\%$ ).

In Study 4.2, a representative sample of 2,807 Dutch participants was recruited via the LISS (Longitudinal Internet Studies of the Social Sciences) panel (Scherpenzeel & Das, 2010). The panel is based on a probability sample of Dutch households drawn from the population register. Panel members are representative of the Dutch population on indicators like gender, age, education, and income.<sup>17</sup> Data from 183 participants (6.52%) who had missing data for at least one question was excluded from analysis, leaving a final sample of 2,624 participants ( $M_{age}$

<sup>17</sup> For more information on the LISS panel, see [lissdata.nl](http://lissdata.nl).

= 52.60,  $SD_{age} = 16.50$ ; 52.52% female, 47.48% male). A sensitivity analysis in G\*Power (Faul et al., 2007) showed that this sample size afforded us 80% power to detect a small difference ( $d = 0.05$ ) when comparing physiognomic beliefs for sociability, morality, and competence (with  $\alpha = 5\%$ ).

**Materials and procedure.** We developed a questionnaire that measures the belief that personality is reflected in facial features. The questionnaire consists of two parts measuring general and specific physiognomic beliefs. First, participants were prompted to imagine seeing the passport photo of a stranger. They were asked to indicate how much they agreed with three statements (e.g., *I can learn something about a person's personality just from looking at his or her face*) on a scale from 1 (*strongly disagree*) to 7 (*strongly agree*). Average scores across the three items constituted our measure of general physiognomic beliefs (Study 4.1: Cronbach's  $\alpha = .74$ , Study 4.2: Cronbach's  $\alpha = .63$ ). We randomized the order in which the three items were presented.

Next, we asked participants how accurately they could judge a person on various characteristics from looking at that person's passport photo. For each item, responses were indicated using a slider from 0 (not at all accurately) to 100 (extremely accurately). Participants responded to twelve items: Three sociability-related traits (*warmth, friendliness, likeability*), three morality-related traits (*trustworthiness, sincerity, honesty*), and three competence-related traits (*competence, intelligence, skillfulness*; Brambilla, Rusconi, Sacchi, & Cherubini, 2011). We also included three additional physically salient characteristics—*gender, age, and attractiveness*. We randomized the order in which the twelve items were presented. Participants' ratings constituted our measure of specific physiognomic beliefs.

In Study 4.1, Dutch participants completed a Dutch version of the scale while non-Dutch participants completed an English version. In Study 4.2, all participants completed the Dutch version of the scale.

## Results

**Study 4.1: Student sample.** In our sample of Dutch students, the average score on general physiognomic belief was just above the scale midpoint ( $M = 4.13$ ,  $SD = 1.23$ ),  $t(377) = 2.03$ ,  $p = .043$ ,  $d = 0.10$ . Around half of all participants (56.88%) believed at least somewhat in physiognomy (i.e., they scored above the midpoint of the scale),  $\chi^2(1) = 6.88$ ,  $p = .009$ .

We calculated average physiognomic belief scores across the nine personality traits to test whether participants think that facial features are more indicative of physically salient characteristics (e.g., a person's gender, age, and attractiveness) than a person's personality. Personality-specific physiognomic beliefs ( $M = 26.68$ ,  $SD = 18.46$ ) were significantly lower than physiognomic beliefs for gender ( $M = 88.30$ ,  $SD = 15.39$ ),  $t(377) = 55.35$ ,  $p < .001$ ,  $d = 2.84$ , attractiveness ( $M = 76.63$ ,  $SD = 18.51$ ),  $t(377) = 44.68$ ,  $p < .001$ ,  $d = 2.30$ , and age ( $M = 66.75$ ,  $SD = 18.11$ ),  $t(377) = 34.06$ ,  $p < .001$ ,  $d = 1.75$ .

We also compared physiognomic beliefs across the three personality dimensions to test whether people think sociability is more reflected in facial features than morality or competence. Physiognomic beliefs were significantly higher for sociability ( $M = 37.12$ ,  $SD = 21.28$ ) compared to morality ( $M = 20.11$ ,  $SD = 18.61$ ),  $t(377) = 20.12$ ,  $p < .001$ ,  $d = 1.03$ , and competence ( $M = 22.81$ ,  $SD = 19.74$ ),  $t(377) = 15.79$ ,  $p < .001$ ,  $d = 0.81$  (see Figure 4.1). Competence-specific physiognomic beliefs were higher than morality-specific physiognomic beliefs, but this difference was less pronounced,  $t(377) = 3.84$ ,  $p < .001$ ,  $d = 0.20$ .

**Study 4.2: Representative sample.** In our representative sample of the Dutch population, the average score on general physiognomic belief was again above the scale midpoint ( $M = 4.17$ ,  $SD = 1.08$ ),  $t(2,623) = 7.96$ ,  $p < .001$ ,  $d = 0.16$ . Around half of all participants (52.10%) believed at least somewhat in physiognomy (i.e., they scored above the midpoint of the scale),  $\chi^2(1) = 4.53$ ,  $p = .033$ .

As in the student sample, personality-specific physiognomic beliefs ( $M = 32.12$ ,  $SD = 18.70$ ) were significantly lower than physiognomic beliefs for gender ( $M = 75.66$ ,  $SD = 19.48$ ),  $t(2,623) = 83.45$ ,  $p < .001$ ,  $d = 1.63$ , age ( $M = 60.52$ ,  $SD = 17.41$ ),  $t(2,623) = 62.52$ ,  $p < .001$ ,  $d = 1.22$ , and attractiveness ( $M = 64.59$ ,  $SD = 20.68$ ),  $t(2,623) = 65.59$ ,  $p < .001$ ,  $d = 1.28$ .

Again, physiognomic beliefs were significantly higher for sociability ( $M = 38.42$ ,  $SD = 21.36$ ) than for morality ( $M = 29.65$ ,  $SD = 20.49$ ),  $t(2,623) = 34.56$ ,  $p < .001$ ,  $d = 0.67$ , or competence ( $M = 28.29$ ,  $SD = 18.76$ ),  $t(2,623) = 34.80$ ,  $p < .001$ ,  $d = 0.68$  (see Figure 4.1). Contrary to findings from Study 4.1, morality-specific physiognomic beliefs were higher than competence-specific physiognomic beliefs, but the difference was again small,  $t(2,623) = 5.77$ ,  $p < .001$ ,  $d = 0.11$ .

We also explored the relationship between physiognomic beliefs and basic demographic indicators. We regressed general physiognomic belief on gender (coded 0 for male and 1 for female), age, income (z-standardized net monthly income), and level of education (six levels, ranging from primary school to university degree). Data from 278 participants (10.59%) who indicated an income of zero and 2 respondents (0.08%) whose reported income was 23.7 and 41.4 standard deviation above the mean were excluded from analysis, leaving a sample of 2,344 participants. Women scored higher on physiognomic belief than men,  $\beta = 0.104$ ,  $SE = 0.048$ ,  $t(2,335) = 2.17$ ,  $p = .030$ , and there was a negative effect of age,  $\beta = -0.004$ ,  $SE = 0.001$ ,  $t(2,335) = 3.07$ ,  $p < .001$ . However, these effects were small: women's physiognomic belief score was 0.097 standard deviations above men's score and a ten-year age difference was associated with a 0.039 standard deviation decrease in belief. There was no effect of income or education (see Supplemental Materials for the full results).

Table 3.1

Descriptive statistics for general physiognomic beliefs across all studies.

Sample	<i>n</i>	$\alpha$	<i>M</i>	<i>SD</i>	% Believers
Study 4.1	378	.74	4.13	1.23	56.88
Study 4.2	2,624	.63	4.17	1.08	52.10
Study 4.3	229	.85	3.92	1.23	47.60
Study 4.4	406	.76	4.14	1.33	54.43
Study 4.5	224	.76	3.97	1.18	48.66

*Note.* General physiognomic beliefs were measured with three items that were rated on a scale from 1 (*strongly disagree*) to 7 (*strongly agree*). “% Believers” indicates the percentage of participants that scored above the midpoint of the scale.

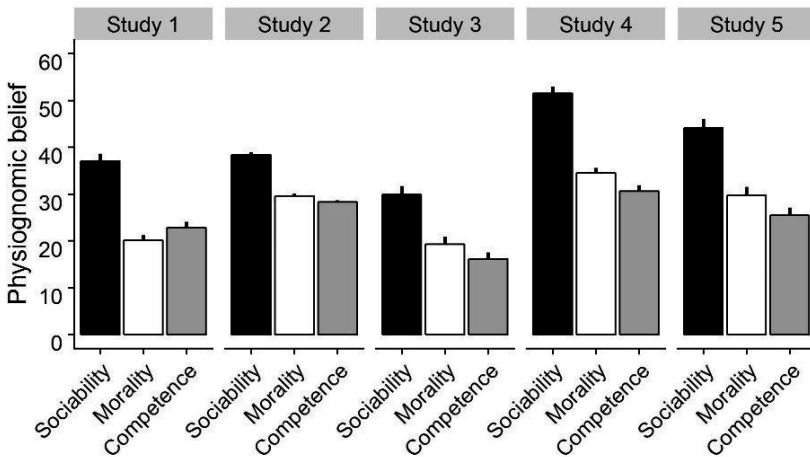


Figure 4.1. Differences in physiognomic beliefs for sociability, morality, and competence across all studies. Error bars represent bootstrapped 95% confidence intervals.

## Discussion

Belief in physiognomy was prevalent in both student and representative samples—over half of all participants at least somewhat

endorsed the belief that personality is reflected in facial features. Physiognomic beliefs were higher among women and older participants, but these differences were small. Moreover, we found no evidence that belief endorsement varied across different levels of education or income. These results suggest that physiognomic beliefs are common across different demographic groups. We also found that people have heterogeneous beliefs about the diagnosticity of facial features for specific characteristics: Participants believed that sociability can be more accurately inferred from facial features than morality or competence. Differences in physiognomic beliefs for morality and competence, however, were small and inconsistent.

### **Study 4.3: Psychological correlates and temporal stability**

Our first studies provided evidence that physiognomic beliefs are relatively common, and that facial appearance is believed to be more indicative of sociability than morality or competence. In Study 4.3, we extended our analysis in two ways. First, we examined the psychological correlates of physiognomic beliefs. That is, we investigated the association of physiognomic beliefs with individual differences in other lay beliefs. Suzuki and colleagues (2017) found that physiognomic beliefs were positively correlated with both entity beliefs (i.e., beliefs in the fixedness and immutability of personality traits; Chiu et al., 1997) and beliefs in biological determinism (i.e., beliefs in the biological determinism of personality traits; Haslam et al., 2004). These findings suggest that physiognomic beliefs are based on the idea that a common factor (e.g., genetic makeup) influences both personality traits and facial appearance (Stirrat & Perrett, 2010). Physiognomic beliefs were also correlated with belief in a just world (Lipkus, 1991), suggesting they may also be rooted in the idea that people “get what they deserve”, with, for example, immoral people having a facial appearance that betrays their immorality to others. We investigated whether the findings by Suzuki



and colleagues (2017) would replicate in a sample of British participants.

Going beyond prior studies, we also investigated whether physiognomic beliefs are related to epistemic motivation. Trait impressions from faces are formed quickly and effortlessly (Stewart et al., 2012; Willis & Todorov, 2006) and this accessibility may make trait impressions intuitively appealing. We therefore expected that endorsement of physiognomic beliefs is more prevalent among people who tend to trust their intuition (i.e., who score high on faith in intuition; Epstein et al., 1996). Relatedly, some people are more prone to override intuitive responses with more analytic and reflective responses (Frederick, 2005; Pennycook, Cheyne, Koehler, & Fugelsang, 2015). We therefore tested whether physiognomic beliefs are negatively related to cognitive reflection. Finally, people vary in their need to form evaluative judgments (Jarvis & Petty, 1996). If people form personality impressions based on facial appearance (which is usually an easily available cue) to satisfy this tendency, endorsement of physiognomic beliefs may justify this behavior. We therefore tested whether physiognomic beliefs are correlated with individual differences in the need to evaluate (Jarvis & Petty, 1996).

Second, we tested the temporal stability of physiognomic beliefs. Participants completed the physiognomic belief scale at two time points with a four-week delay

## Methods

**Participants.** In the study of Suzuki and colleagues (2017), correlations ranging from  $r = .185$  to  $r = .445$  ( $n = 1,396$ ) were reported. We therefore aimed to recruit 227 participants, which affords 80% power to detect a correlation of  $r = .185$  (with  $\alpha = 5\%$ ). We recruited 310 British Prolific workers to complete the study in exchange for £1.25 each. In line with our preregistration, data from 79 participants (25.48%) who

failed an attention check question at the end of the study and from 2 participants (0.87%) who indicated poor or basic English proficiency were excluded from analysis, leaving a final sample of 229 participants ( $M_{age} = 35.62$ ,  $SD_{age} = 11.86$ ; 60.26% female, 39.30% male, 0.44% other).

To measure the temporal stability of physiognomic beliefs, we re-contacted participants after four weeks. An a priori power analysis showed that a sample size of 84 participants is required to detect a medium-sized correlation ( $r = .300$ ) between physiognomic belief scores at both time points with 80% power (and  $\alpha = 5\%$ ). We re-contacted a total of 200 British Prolific workers to complete the second part of the study in exchange for £0.50 each. In line with our preregistration, data from 55 participants (27.50%) who failed an attention check question at the end of the study and from 16 participants (11.03%) whose responses could not be matched with data from part one were excluded from analysis, leaving a final sample of 129 participants ( $M_{age} = 38.26$ ,  $SD_{age} = 12.16$ ; 60.47% female, 38.76% male, 0.78% other).

**Materials and procedure.** Physiognomic beliefs were measured as described in Study 4.1.

Belief in the entity theory of personality was measured with eight items (e.g., "The kind of person someone is, is something basic about them, and it can't be changed very much") adapted from Levy and colleagues (1998; Study 5). Participants indicated how much they agreed with each statement on a scale ranging from 1 (*strongly disagree*) to 6 (*strongly agree*).

Following previous investigations (Haslam et al., 2004; Suzuki et al., 2017), we measured belief in biological determinism of personality traits by showing participants a list of nine personality traits (the same nine personality traits representing evaluations of sociability, morality, and competence that are used for the physiognomic belief scale). We asked them to rate how much each trait is based on biological nature

(genes, brain structure, etc.) on a scale ranging from 0 (*not based on biological nature*) to 100 (*based on biological nature*).

Belief in a just world was measured with seven items (e.g., “I feel that people get what they deserve”) adapted from Lipkus (1991). Participants indicated how much they agreed with each statement on a scale ranging from 1 (*strongly disagree*) to 6 (*strongly agree*).

Faith in intuition was measured with twelve items (e.g., “I trust my initial feelings about people”) adapted from Epstein and colleagues (1996). Participants indicated how much they agreed with each statement on a scale ranging from 1 (*completely false*) to 5 (*completely true*).

Cognitive reflection was measured with the seven-item cognitive reflection test (CRT; e.g., “If you are in a race and you pass the person in second place, what place are you in?”) adapted from Thomson and Oppenheimer (2016). The CRT measures the tendency to override an intuitive, but incorrect answer with a more reflective and correct one. Participants indicated their responses in a free form text box. The test was scored by adding up the number of items that were answered correctly.

Need to evaluate was measured with sixteen items (e.g., “I form opinions about everything”) adapted from Jarvis and Petty (1996). Participants indicated to what extent each item was characteristic of them on a scale ranging from 1 (*extremely uncharacteristic*) to 5 (*extremely characteristic*).

Participants completed the six measures, and the items within each measure, in a random order. We randomized whether participants completed the physiognomic belief scale before or after the other measures. For the second part of the study which was conducted four weeks later, participants completed the physiognomic belief scale a second time.

## Results

The average score on general physiognomic belief was just below the midpoint of our scale ( $M = 3.92$ ,  $SD = 1.23$ ),  $t(228) = 0.98$ ,  $p = .33$ ,  $d = 0.06$  (see Table 4.1). Around half of all participants (47.60%) believed at least somewhat in physiognomy (i.e., they scored above the midpoint of the scale),  $\chi^2(1) = 0.44$ ,  $p = .51$ . All scales showed acceptable to excellent internal consistency ( $.72 < \alpha < .92$ ; see Table 4.2).

**Psychological correlates.** First, we examined the relationship between physiognomic beliefs and other lay beliefs. Participants who scored higher on physiognomic beliefs also scored higher on belief in the biological determinism of personality traits,  $r(227) = .172$ ,  $p = .009$ , and belief in a just world,  $r(227) = .154$ ,  $p = .020$ .<sup>18</sup> We did not find that physiognomic beliefs were related to belief in the entity theory of personality,  $r(227) = .093$ ,  $p = .16$ .

Next, we examined how physiognomic beliefs relate to epistemic motivation. Participants who scored higher on physiognomic beliefs also scored higher on faith in intuition,  $r(227) = .409$ ,  $p < .001$ .<sup>19</sup> We did not find any evidence that physiognomic beliefs were related to scores on the cognitive reflection test,  $r(227) = -.048$ ,  $p = .46$ . The correlation with need to evaluate was positive, but only marginally significant,  $r(227) = .125$ ,  $p = .058$ .

We also explored which of the measures were uniquely related to physiognomic beliefs, by regressing physiognomic beliefs on all six measures, which were z-standardized prior to analysis. This yielded a positive effect of faith in intuition,  $\beta = 0.484$ ,  $SE = 0.087$ ,  $t(222) = 5.59$ ,  $p$

---

<sup>18</sup> The correlation with belief in a just world was no longer significant when correcting for multiple comparisons (see Supplemental Materials).

<sup>19</sup> The faith in intuition scale includes two items that directly refer to the accuracy of appearance-based impressions (“My initial impressions of people are almost always right” and “I believe I can judge character pretty well from a person’s appearance”). Physiognomic beliefs were still correlated with faith in intuition when these two items were omitted,  $r = .345$ ,  $p < .001$ .

< .001, but no significant effects of entity beliefs,  $\beta = 0.078$ ,  $SE = 0.77$ ,  $t(222) = 1.01$ ,  $p = .31$ , beliefs in biological determinism,  $\beta = 0.045$ ,  $SE = 0.080$ ,  $t(222) = 0.57$ ,  $p = .57$ , belief in a just world,  $\beta = 0.091$ ,  $SE = 0.78$ ,  $t(222) = 1.17$ ,  $p = .24$ , cognitive reflection,  $\beta = 0.024$ ,  $SE = 0.076$ ,  $t(222) = 0.31$ ,  $p = .75$ , or need to evaluate,  $\beta = -0.024$ ,  $SE = 0.083$ ,  $t(222) = 0.29$ ,  $p = .77$ .

Table 4.2

Descriptive statistics and correlations for all personality variables.

Measure	M	SD	$\alpha$	Correlation						
				1	2	3	4	5	6	
1. PB	3.92	1.23	.85	—	—	—	—	—	—	—
2. BET	3.30	0.90	.92	.093	—	—	—	—	—	—
3. BBD	43.61	18.57	.88	.172 ***	.229 ***	—	—	—	—	—
4. BJW	2.98	0.81	.87	.154 *	.017	.104	—	—	—	—
5. FI	3.56	0.60	.85	.409 ***	.054	.253 ***	.183 **	—	—	—
6. CRT	3.73	1.91	.72	-.048	.019	-.037	.016	-.177 **	—	—
7. NE	3.48	0.60	.82	.125 †	.079	.116 †	-.125 †	.370 ***	-.009	—

Note. PB = Physiognomic belief, BET = Belief in the entity theory of personality, BBD = Belief in the biological determinism of personality traits, BJW = Belief in a just world, FI = Faith in intuition, CRT = Cognitive reflection, NE = Need to evaluate.

†  $p < .10$ , \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$

**Temporal stability.** The average score on general physiognomic belief when measured four weeks later was just below the midpoint of our scale ( $M = 3.99$ ,  $SD = 1.18$ ),  $t(128) = 0.12$ ,  $p = .91$ ,  $d = 0.01$ . Around half of all participants (48.84%) believed at least somewhat in physiognomy (i.e., scored above the midpoint of the scale),  $\chi^2(1) = 0.03$ ,  $p = .86$ . We found a strong correlation between general physiognomic belief scores at both time points,  $r(127) = .644$ ,  $p < .001$ , and between specific physiognomic beliefs (averaged across all nine personality traits) at both time points,  $r(127) = .609$ ,  $p < .001$ .<sup>20</sup>

## Discussion

The current study provided new insights into who believes in physiognomy. In line with Suzuki and colleagues (2017), we found that people who score high on physiognomic belief were more likely to endorse the belief that personality traits are determined by biological factors. This suggests that physiognomic beliefs may be rooted in the idea that biological factors (e.g., genetic makeup) determine both personality and facial appearance. We did not find evidence that physiognomic beliefs were related to entity beliefs, and the correlation with belief in a just world was weak.

We did find a correlation between physiognomic beliefs and a measure of intuitive thinking style. Trait inferences from faces are formed spontaneously, quickly, and effortlessly and can serve as input for intuitive judgments or decisions (Freeman & Johnson, 2016; Jaeger, Evans, Stel, et al., 2019a). Thus, the intuitive accessibility of trait inferences may make them particularly appealing for individuals who tend to follow their intuitions. In line with this view, people who scored high on faith in intuition endorsed physiognomic beliefs more strongly. Moreover, predicting physiognomic beliefs with all individual

---

<sup>20</sup> Correlation coefficients ranged from .43 for age-specific belief to .72 for competence-specific belief.

differences measured here showed only a unique effect of intuitive thinking style. At the same time, we did not find a correlation with scores on the cognitive reflection test. This result suggests that reliance on physiognomic beliefs stems from the preference to rely on intuition, rather than an inability to override intuitive response.

#### **Study 4.4: Confidence, accuracy, and meta-accuracy**

In Study 4.4, we tested how individual differences in physiognomic beliefs are related to actual and predicted accuracy of trait impressions from faces. We aimed to address three questions. First, we examined whether physiognomic beliefs are related to confidence in the accuracy of trait impressions. Research shows that people are relatively confident in the accuracy of their physiognomic judgments, in spite of their generally low actual accuracy (Ames et al., 2010; Biesanz et al., 2011). We propose that this confidence may result from subjective beliefs that faces are a good indicator of personality.

Second, we examined whether physiognomic beliefs are related to the actual accuracy of trait impressions. People might endorse physiognomic beliefs because their physiognomic judgments are indeed more accurate. Third, we examined whether superior judgment accuracy explains the relationship between physiognomic beliefs and confidence. People who believe in physiognomy may be justifiably more confident in their trait impressions because their impressions are more accurate. Alternatively, physiognomic beliefs may influence confidence irrespective of judgment accuracy. This would imply that increased confidence by people who believe in physiognomy is not normatively justified. To test these competing accounts, we examined whether the relationship between physiognomic beliefs and confidence is mediated by judgment accuracy.

To answer these questions, we replicated a previous study on the accuracy of trait impressions from faces (C. Lin et al., 2018). In this study,



participants rated the corruptibility of government officials who had a clean record or who had been found guilty of political corruption. Lin and colleagues (2018) found that accuracy in corruptibility detection based on facial photographs was significantly above chance and people varied in how accurate their judgments were. Here, we gathered corruptibility ratings of the same photo stimuli to measure accuracy in corruptibility judgments. We measured confidence by asking participants to estimate how many individuals they would classify correctly. Participants also completed the physiognomic belief scale. Given that participants specifically judged corruptibility, which is conceptually similar to trustworthiness (C. Lin et al., 2018), we analyzed their trustworthiness-specific physiognomic beliefs.

## Methods

**Participants.** An a priori power analysis showed that a sample size of 193 participants is required to detect a small-to-medium-sized correlation between physiognomic beliefs and confidence in trait judgments ( $r = .20$ ) with 80% power (and  $\alpha = 5\%$ ). We therefore aimed to recruit at least 193 participants, with the final sample size being determined by how many students participated in the study in two weeks. In total, we recruited 512 first-year psychology students from a Dutch university who completed the study in return for partial course credit. In line with our preregistration, data from 101 participants (19.73%) who indicated poor or basic English proficiency, from 3 participants (0.01%) who always indicate the same rating (corruptible or not corruptible) across all trials, and from 1 participant (0.002%) whose response time was faster than 100 milliseconds on at least 10% of all trials were excluded, leaving a final sample of 406 participants ( $M_{age} = 20.01$ ,  $SD_{age} = 2.17$ ; 80.54% female, 19.21% male, 0.25% other).

**Materials and procedure.** We used an image set of 72 US government officials created by Lin and colleagues (2018, Study 1). Half

of the politicians were convicted of political corruption, whereas the other half had clean records. The images were obtained from personal websites, news articles, or Wikipedia. The faces were converted to gray-scale, cropped to a uniform size, and shown against a uniform background.

Participants saw the 72 images in a randomized order and were asked to indicate whether they thought the politician in each photo was *corruptible* (i.e., *untrustworthy, dishonest, selfish*) or not (response options were “yes” or “no”). Next, participants were asked whether they had recognized any of the individuals and, in case they answered affirmatively, whom they had recognized. None correctly identified any politicians. We measured confidence in the accuracy of impressions by asking participants to rate how often they think they made the right judgment on a scale from 0% of the time to 100% of the time. Participants were reminded that, given the two-alternative forced choice design, 50% accuracy would be expected by chance. Finally, participants completed the physiognomic belief scale (Cronbach’s  $\alpha = .76$ ). All participants completed the study in English.

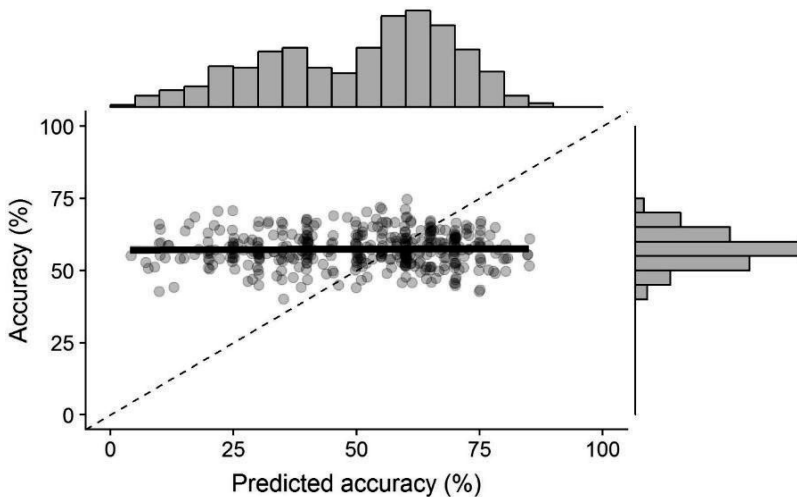
## Results

The average score on general physiognomic belief was just above the midpoint of our scale ( $M = 4.14$ ,  $SD = 1.13$ ),  $t(405) = 2.53$ ,  $p = .012$ ,  $d = 0.13$  (see Table 4.1). Around half of all participants (54.43%) believed at least somewhat in physiognomy (i.e., they scored above the midpoint of the scale),  $\chi^2(1) = 3.02$ ,  $p = .082$ .

**Accuracy and meta-accuracy.** First, we investigated whether face-based corruptibility judgments were accurate. We examined individual-level accuracy by testing whether the percentage of times participants made a correct judgment (i.e., judging a corrupt politician as corruptible or a politician with a clean record as incorruptible) was higher than 50% with a one-sample one-tailed  $t$ -test. The average

individual-level accuracy was 57.35% ( $SD = 5.93$ ), which was significantly higher than chance,  $t(405) = 193.16$ ,  $p < .001$ ,  $d = 9.59$ . A total of 356 participants (87.58%) identified more than 50% of politicians correctly. Thus, participants' impressions were somewhat accurate.

Were participants aware of the accuracy of their judgments? On average, participants expected their judgments to be correct on 50.39% ( $SD = 18.60$ ) of all trials, which was significantly below their actual accuracy,  $t(405) = 7.54$ ,  $p < .001$ ,  $d = 0.37$ . The correlation between individual accuracy levels and predicted accuracy levels was not significant and close to zero,  $r(404) = .018$ ,  $p = .72$  (see Figure 4.2). In sum, we found no evidence for meta-accuracy and on average, participants underestimated their accuracy.



*Figure 4.2.* Correlation between predicted and actual accuracy in corruptibility judgments. The diagonal line represents perfect meta-accuracy (i.e., predicted accuracy corresponds to actual accuracy). Data points to the left of the line represent participants who underestimated accuracy and data points to the right of the line represent participants who overestimated accuracy. The histograms in the margins show the distributions of predicted and actual accuracy.

**Physiognomic belief.** Our next set of analyses investigated the relationships between physiognomic beliefs, individual-level judgment accuracy, and confidence in judgment accuracy. We expected participants scoring higher on physiognomic beliefs to be more confident in the accuracy of their judgments. In fact, endorsement of physiognomic beliefs was positively correlated with predicted judgment accuracy,  $r(404) = .285, p < .001$  (see Figure 4.3, left panel).

Next, we examined whether people who scored higher on physiognomic beliefs were actually more accurate in their judgments. There was a positive correlation between physiognomic beliefs and judgment accuracy,  $r(404) = .167, p < .001$  (see Figure 4.3, right panel). Was increased confidence by people scoring high on physiognomic beliefs justified? If better judgment accuracy explains the positive relationship between physiognomic beliefs and confidence in accuracy, then we would expect that accuracy mediates the link between physiognomic beliefs and confidence. However, there was no significant indirect effect of physiognomic belief on confidence via accuracy,  $\beta = -0.050, p = .49, 95\% \text{ CI } [-0.452, 0.200]$ .

We also regressed confidence and accuracy on physiognomic beliefs (in separate models) and compared the strength of the effects with a  $z$ -test. This showed that the effect of physiognomic beliefs on confidence ( $\beta = 5.135, SE = 0.890, t(404) = 5.97, p < .001$ ) was significantly stronger than the effect of physiognomic beliefs on accuracy ( $\beta = 0.992, SE = 0.292, t(404) = 3.40, p < .001$ ),  $z = 4.62, p < .001$ . Participants who scored one standard deviation higher on physiognomic beliefs expected to be 5.32 percentage points more accurate, but were only 0.99 percentage points more accurate. In sum, the increased confidence of people scoring high on physiognomic beliefs was not warranted given their actual judgment accuracy. The increase in confidence was disproportionately larger compared to the increase in actual accuracy.

Finally, we explored whether physiognomic beliefs were related to *overconfidence* in face-based impressions. We subtracted actual accuracy from predicted accuracy to create a variable indicating by how much participants overestimated their accuracy. Physiognomic beliefs were positively correlated with overestimation,  $r(404) = .222, p < .001$ . Thus, people scoring higher on physiognomic beliefs were more likely to be overconfident in their judgment accuracy.

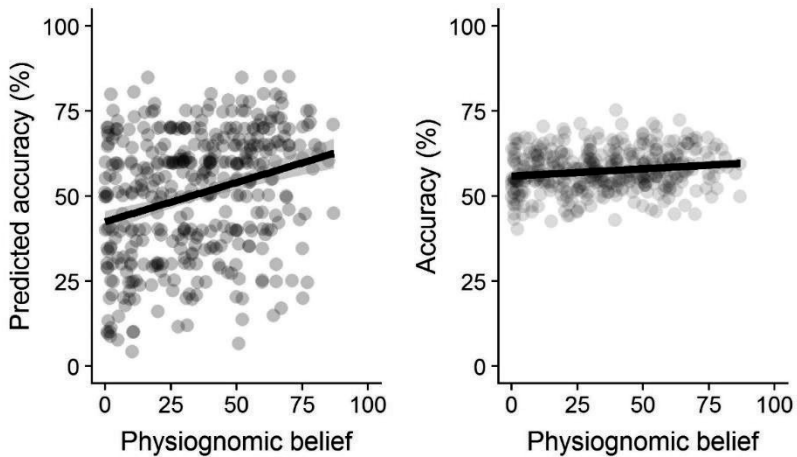


Figure 4.3. Correlations between physiognomic beliefs and confidence (i.e., predicted accuracy of corruptibility judgments; left), and actual accuracy of judgments (right).

## Discussion

We found that people who believe in physiognomy were more accurate in judging the corruptibility of government officials based on facial photographs. Physiognomic beliefs were also related to confidence in judgments: Participants who more strongly endorsed physiognomic beliefs expected their corruptibility judgments of politicians to be more accurate. Crucially, we found that the increase in confidence was

disproportionately larger than the increase in judgment accuracy. In other words, physiognomic beliefs were related to overconfidence in trait impressions. Moreover, the correlation between accuracy and confidence was non-significant and close to zero. This shows that the increased confidence of people endorsing physiognomic beliefs is not normatively justifiable, as it is not based on their actual judgment accuracy.

### **Study 4.5: Reliance on judgments**

We proposed that physiognomic beliefs may help explain why people overrely on face-based personality judgments (Olivola, Funk, et al., 2014). Results from Study 4.4 were in line with this view, showing that people who scored higher on physiognomic beliefs were too confident in the accuracy of their judgments. In Study 4.5, we extended these findings by examining whether people who score higher on physiognomic beliefs also rely more on face-based trait impressions in social decision-making. More specifically, we tested whether physiognomic beliefs relate to reliance on trait impressions when people could also rely on more valid information instead.

We tested this hypothesis in the context of the trust game (Berg, Dickhaut, & McCabe, 1995; Snijders & Keren, 1999). This interaction between two players, a *trustor* and a *trustee*, reflects the essential structure of trust-based social exchange. Participants played a series of trust games and, in each round, they saw a photo of their interaction partners. This allowed us to measure how much participants relied on the perceived facial trustworthiness of their interaction partner when making trust decisions. In real life, people can often rely on a variety of cues when making decisions and the persistent influence of facial trustworthiness is particularly surprising in cases where people could rely on more valid cues instead (Olivola et al., 2018; Rezlescu et al., 2012). We therefore varied a second cue that participants could rely on:

the trustee's temptation to choose betrayal, which actually predicts the likelihood of reciprocation (Evans & Krueger, 2014).

We expected that individual differences in physiognomic beliefs would be related to reliance on facial trustworthiness, but not to reliance on temptation. That is, we predicted that people who believe in physiognomy would rely more on their face-based trustworthiness judgments, but not necessarily more on other cues unrelated to face-based judgments.

## Methods

**Participants.** We recruited 243 first-year psychology students from a Dutch university who completed the study in return for partial course credit. Due to a technical error, some face images were not displayed for 19 participants and we therefore excluded their data from analysis, leaving a final sample of 224 participants ( $M_{age} = 20.45$ ,  $SD_{age} = 2.41$ ; 75.45% female, 24.11% male, 0.45% other).<sup>21</sup> The majority of participants were Dutch (43.75%) or German (30.36%). Sample size was determined by how many students participated in the study within two weeks. A sensitivity analysis in G\*Power (Faul et al., 2007) showed that this sample size afforded us 80% power to detect a small correlation ( $r = .186$ ) between physiognomic beliefs and reliance on facial trustworthiness (with  $\alpha = 5\%$ ).

**Materials and procedure.** The experiment was administered online. Participants first learned about and then played a series of 24 trust games in the role of the trustor. On each trial, participants saw a photo of their supposed interaction partner next to the decision tree. We selected 12 photos of Caucasian Dutch adults (six females and six males) with a forward gaze from the Radboud Faces Database (Langner et al.,

---

<sup>21</sup> Participants were drawn from the same subject pool as participants in Study 4.1. We could not check for potential overlap between samples as no identifying information was collected in either study.

2010). To introduce variance in perceived trustworthiness, half of the selected faces displayed a neutral expression and half a happy facial expression (Evans & van de Calseyde, 2017; Krumhuber et al., 2007).

We also varied the trustee's temptation, the economic incentive to choose betrayal. We defined temptation as the difference between the trustee's gain in case of betrayal ( $T$ ) and reciprocation ( $R_2$ ) divided by the value of betrayal:  $(T - R_2)/T$  (Evans & Krueger, 2014). Each photo was shown twice, once when temptation was low (0.33) and once when temptation was high (0.60). These values correspond to a 50% (low temptation) and 150% (high temptation) increase in payoffs for the trustee in case betrayal is chosen over reciprocation.

After indicating their trust decisions, participants were shown each face again and asked to rate how trustworthy they think the person in the photo is on a scale from 1 (*not at all trustworthy*) to 9 (*extremely trustworthy*). We used the average trustworthiness rating of each face across all participants as our measure of facial trustworthiness. Prior to analysis, we rescaled the cue variables (i.e., facial trustworthiness and temptation) to range from -0.5 to 0.5. Thus, for the two cues, a one-unit increase denotes a change from the lowest average trustworthiness rating to the highest and a change from low to high temptation. Finally, participants completed the physiognomic beliefs scale ( $\alpha = .76$ ). All participants completed the study in English.

## Results

The average score on general physiognomic belief was below the midpoint of our scale ( $M = 3.97, SD = 1.18$ ),  $t(223) = 0.32, p = .75, d = 0.02$  (see Table 4.1). Around half of all participants (48.66%) believed at least somewhat in physiognomy (i.e., they scored above the midpoint of the scale),  $\chi^2(1) = 0.11, p = .74$ . Average trustworthiness ratings of the faces ranged, from 4.04 to 6.61 on our 9-point scale ( $M = 5.21, SD = 0.88$ ). We computed intraclass correlation coefficients (ICCs) to estimate



consensus in ratings across participants (Shrout & Fleiss, 1979). Participants showed significant consensus in their ratings,  $ICC(2, 1) = .261, p < .001, 95\% \text{ CI } [.132, .597]$ . The average trust rate across all trials was 46.09% and participants took on average 5.43 seconds ( $SD = 3.50$ ) to make a decision. Eight participants (3.57%) never trusted whereas six participants (2.68%) always trusted.

**Reliance on facial trustworthiness and temptation.** First, we checked whether participants relied on facial trustworthiness and temptation when making trust decisions. We estimated a multilevel regression model with random intercepts and slopes per participant in which we regressed trust decisions (coded 0 for no trust and 1 for trust) on facial trustworthiness and temptation. This revealed a positive effect of facial trustworthiness,  $\beta = 1.608, SE = 0.158, z = 10.17, p < .001, OR = 4.99$ . The individual with the lowest trustworthiness rating was trusted 29.96% of the time while the individual with the highest trustworthiness rating was trusted 62.11% of the time. There was also a negative effect of temptation,  $\beta = -0.463, SE = 0.090, z = 5.12, p < .001, OR = 0.63$ . Participants trusted 50.55% of the time when temptation was low and 42.07% of the time when temptation was high. Thus, participants relied on both cues when making trust decisions.

**Physiognomic beliefs and cue reliance.** Next, we tested our main hypothesis that physiognomic beliefs are correlated with reliance on facial trustworthiness. We extracted the participant-specific effects of temptation and facial trustworthiness (i.e., the random slopes) from our multilevel regression models as an indicator of how much each participant relied on the two cues. Both cue reliance variables were z-standardized. Results showed a positive correlation between physiognomic beliefs and reliance on facial trustworthiness,  $r(222) = .142, p = .034$ , but no significant correlation with reliance on temptation,  $r(222) = .014, p = .83$ . Exploratory analyses (see Supplemental Materials) showed that there were also no significant correlations with reliance on

other facial features (i.e., facial expression, gender, attractiveness). Moreover, regressing physiognomic beliefs on all cue reliance variables showed only a significant effect of reliance on facial trustworthiness. Thus, participants who scored higher on physiognomic beliefs relied more on facial trustworthiness, but not more on other cues, when making trust decisions.

### **Discussion**

In line with previous studies, we found that participants relied on the facial appearance of their interaction partners when deciding whom to trust (Jaeger, Evans, Stel, et al., 2019a; van't Wout & Sanfey, 2008). More importantly though, we found that *how much* participants relied on trustworthiness impressions from face was related to their endorsement of physiognomic beliefs. Specifically, belief that trustworthiness is reflected in facial features predicted to what extent trust decisions were driven by the facial appearance of interaction partners. Physiognomic beliefs were only correlated with reliance on the perceived trustworthiness of interaction partners and not with reliance on other cues such as the interaction partner's facial expression, attractiveness, gender, or their temptation (i.e., their incentive to betray trust).

### **General discussion**

The goal of the current investigation was to provide insights into the belief that personality is reflected in facial appearance (i.e., physiognomic beliefs). The accuracy of face-based personality inferences has been studied since the time of Ancient Greece (Aristotle, trans. 1936) and remains a subject of contentious debate (Bonnefon et al., 2015; Todorov, Funk, et al., 2015). Here, we argue that irrespective of their actual accuracy, people hold beliefs about the diagnosticity of facial features for inferring personality. Crucially, individual differences in the endorsement of physiognomic beliefs may help explain why and under

what conditions people are confident in their personality impressions, or rely on them to make decisions. To test these hypotheses, we introduced a scale to measure physiognomic beliefs and examined their prevalence, structure, and correlates.

First, our results suggest that belief in physiognomy is relatively widespread. Across all studies, around half of all participants at least somewhat endorsed the belief that personality is reflected in facial features. Physiognomic beliefs were prevalent among psychology students at a Dutch university (Studies 4.1, 4.4, and 4.5), in the general Dutch population (Study 4.2), and among participants from the United Kingdom recruited through a crowdsourcing platform (Study 4.3). A closer analysis of our representative sample of the Dutch population showed that physiognomic beliefs were slightly more prevalent among women and younger participants, but these differences were small. We found no significant differences across different levels of education or income. Together, these results suggest that belief in physiognomy is common across different demographic groups.

Second, people hold heterogeneous beliefs about the manifestation of different traits in faces. To map physiognomic beliefs across a variety of characteristics, we measured beliefs in three fundamental dimensions underlying person perception: sociability, morality, and competence (Brambilla et al., 2011). Participants believed that sociability is more reflected in faces than morality or competence and this pattern replicated in all five studies (see Figure 4.1). Differences in physiognomic beliefs for morality and competence were small and inconsistent across studies.

Third, physiognomic beliefs are related to an intuitive thinking style. Personality impressions from faces are formed spontaneously, quickly, and effortlessly (Klapper et al., 2016; Stewart et al., 2012; Willis & Todorov, 2006). We therefore hypothesized that people who tend to trust their intuitions would be more likely to endorse physiognomic

beliefs. Results from Study 4.3 supported this prediction. Physiognomic beliefs were also correlated with beliefs in the biological determinism of personality traits (replicating results of Suzuki et al., 2017), but this relationship was less pronounced and non-significant when controlling for faith in intuition. These results suggest that physiognomic beliefs may be rooted in the quick and efficient processing of faces which makes trait inferences from faces intuitively accessible.

Fourth, physiognomic beliefs shape how impressions from faces are formed and subsequently used in decision-making. People who more strongly endorsed physiognomic beliefs were more confident in the accuracy of their face-based personality judgments (Study 4.4) and relied more on face-based personality judgments when deciding whom to trust (Study 4.5). In fact, trustworthiness-specific physiognomic beliefs were correlated with reliance on an interaction partner's perceived trustworthiness, but not with reliance on other facial or non-facial cues (e.g., the interaction partner's attractiveness or their economic incentive to betray trust).

Finally, the increased weighing of personality inferences by people scoring higher on physiognomic beliefs is not due to superior judgment accuracy. We asked whether people who endorse physiognomic beliefs were justifiably more confident because of their superior judgment accuracy (Study 4.4). To this end, we replicated a recent study by Lin and colleagues (2017) that demonstrated accuracy in corruptibility judgments of politicians based on face images. We found that corruptibility judgments by people scoring higher on physiognomic beliefs were indeed slightly more accurate, suggesting that individual differences in physiognomic beliefs may reflect superior accuracy in inferring personality from faces. However, mediation analyses showed that judgment accuracy did not account for the link between physiognomic beliefs and confidence. The absolute increase in confidence was also disproportionately larger than the increase in

accuracy: A one standard deviation increase in physiognomic beliefs was related to a one percentage point increase in accuracy, but a five percentage point increase in predicted accuracy. In other words, the observed advantage in accuracy did not justify the increase in confidence and, on average, people who endorsed physiognomic beliefs were more likely to overestimate the accuracy of their judgments.

### **Lay personality theory**

People hold beliefs about the basis (Haslam et al., 2004), malleability (Chiu et al., 1997), structure (Stolier, Hehman, Keller, et al., 2018), and expression (Mehl et al., 2006) of personality traits. We add to this work by showing that people also hold beliefs about the manifestation of personality traits in facial appearance. While endorsement of this belief varied across individuals, there was considerable consistency in belief structure across different personality dimensions. Across all five studies, physiognomic beliefs were strongest for sociability compared to morality and competence. People vary in their absolute belief in physiognomy, but beliefs about the relative expression of different personality dimensions in faces is, to a large extent, shared.

Sociability and morality are often subsumed under the label of warmth (S. T. Fiske et al., 2002) or communion (Abele & Wojciszke, 2007). However, judgments of sociability and morality show several important differences and researchers have argued that they should be treated as separate dimensions of person evaluation (Brambilla et al., 2011; Goodwin, 2015; Landy, Piazza, & Goodwin, 2016). For instance, morality information more strongly determines the formation (Goodwin et al., 2014; Leach et al., 2007) and updating (Brambilla, Carraro, Castelli, & Sacchi, 2019) of impressions. We showed that sociability and morality also displayed divergent patterns in the domain of lay personality theory.

## **Social perception**

Dominant theories on impression formation from faces have mostly focused on how different facial features elicit trait inferences (Oosterhof & Todorov, 2008; Zebrowitz, 2017), which has produced a long list of cues that people use to infer personality from faces (e.g., Jaeger, Wagemans, Evans, & van Beest, 2018; Said, Sebe, & Todorov, 2009; Sutherland, Young, & Rhodes, 2016). This approach reflects the view that social perception is mostly a reflexive, stimulus-driven processes in which the presence of certain facial cues automatically trigger personality inferences (Engell et al., 2007; van't Wout & Sanfey, 2008; Winston et al., 2002). However, recent studies have highlighted that there are many top-down processes that influence social perception (Brambilla et al., 2018; Freeman & Johnson, 2016). For example, beliefs about the extent to which personality traits correlate intra-personally (e.g., whether trustworthy people tend to be sociable) influence the correlation between personality trait impressions from faces (e.g., the overlap in impression of trustworthiness and sociability; Stolier, Hehman, & Freeman, 2018; Stolier, Hehman, Keller, et al., 2018). In a similar vein, our results showed that beliefs about the manifestation of personality traits in faces influences confidence in and reliance on trait impressions. Thus, the processing of personality trait information from faces is moderated by individual differences in lay personality theory.

Despite their poor predictive validity, trait inferences from faces influence a wide range of consequential decisions, such as criminal sentencing, voting, and personnel selection (Olivola, Funk, et al., 2014) and people still rely on facial appearance when better information is available (Jaeger, Evans, Stel, et al., 2019a; Olivola & Todorov, 2010b; Rezlescu et al., 2012). The influence of facial stereotypes can lead to worse outcomes for decision-makers, and to systematic discrimination against people of a certain appearance. Although the effects of this bias are well-documented, little is known about the mechanisms underlying

it. Yet, addressing the bias—for example, by implementing interventions that reduce reliance on facial appearance—requires knowledge about its cognitive underpinnings (Forscher et al., 2019; T. D. Wilson & Brekke, 1994). The current results suggest that widespread influence of facial stereotypes may be explained by lay beliefs in the diagnosticity of facial appearance as an indicator of personality. As a consequence, people are overconfident in the accuracy of their face-based trait inferences and rely on them too much when making decisions. Future studies could test whether changing physiognomic beliefs by educating people about the low predictive validity of their impressions reduces reliance on facial appearance.

### **Limitations and future directions**

Our findings point to a widespread belief in the core tenet of physiognomy (Aristotle, 1936; Lavater, 1775) that personality traits are reflected in facial morphology. But did participants actually envision a resting, non-expressive face or is their belief rooted in the perceived informational value of dynamic features of faces, such as emotion expressions? This distinction may ultimately be inconsequential because people readily perceive emotion expressions in resting faces (Adams et al., 2012; Said et al., 2009). We tried to ensure that participants imagine a resting face by prompting them to imagine seeing the passport photo of a stranger. In both the Netherlands and the United Kingdom, the countries of origin of most participants, people are required to maintain a neutral expression in passport photos. Future studies could investigate the role of emotional expressions, for example, by comparing confidence in personality judgments of neutral and expressive faces.

Future research could also explore how physiognomic beliefs vary across different social groups. The majority of our studies relied on student samples from a Dutch university, which constrains the generalizability of our findings. We did find that belief in physiognomy

was common in a sample of British participants and in a representative sample of the Dutch population. Moreover, Study 4.2 showed that belief in physiognomy varied little across different demographic indicators (i.e., gender, age, education, and income). Nonetheless, research with participants from more diverse cultural backgrounds is needed to investigate who believes in physiognomy. In addition, future studies could leverage the rich and openly accessible data of the LISS panel—from which we recruited a representative sample of Dutch participants for Study 4.2—to map how different psychological or socio-economic variables relate to physiognomic beliefs.

Another question that remains unanswered is whether individual differences in physiognomic beliefs predict which cues people rely on to form impressions. For example, trustworthiness impressions are based on a wide variety of interrelated facial cues, such as width-to-height ratio (Stirrat & Perrett, 2010), resemblance to emotion expressions (Said et al., 2009), and sexual dimorphism (Gladstone & O'Connor, 2014). Laypeople seem to agree that trustworthiness is reflected in facial appearance, but do they use the same cues to infer trustworthiness? If physiognomic beliefs are rooted in a greater ability to infer personality from faces—and Study 4.4 provided some support for this—then people who believe in physiognomy should not only rely on similar cues, but they should rely on the cues that are actually valid indicators of the trait in question. These questions could be addressed in a lens model framework (Brunswik, 1956), by testing whether physiognomic beliefs are related to cue utilization (i.e., how much people rely on different cues) and cue validities (i.e., how well these cues predict the criterion being judged).

### **Conclusion**

We showed that people hold lay beliefs about the manifestation of personality traits in facial appearance (i.e., physiognomic beliefs). While



people differ in their absolute endorsement of physiognomy, beliefs about how much different personality dimensions are reflected in facial features are largely shared. We also find that individual differences in physiognomic beliefs are related to various aspects of social perception. People who score high on physiognomic beliefs are more confident in the accuracy of their personality judgments (but this cannot be explained by their superior judgment accuracy) and they rely more on their personality judgments when making decisions. In sum, our results show that physiognomic beliefs are widespread and associated with a range of social-cognitive processes and behaviors.



# Chapter 5

The bounds of physiognomy: Lay beliefs in the manifestation of personality traits in facial features

Based on:

Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2019). *The bounds of physiognomy: Lay beliefs in the manifestation of personality traits in faces*. Manuscript submitted for publication.

All data, preregistration documents, and analysis scripts are available at the Open Science Framework (<https://osf.io/cbsmw/>).

## Abstract

Even though personality trait impressions from faces are generally inaccurate, beliefs in the diagnostic value of facial appearance for judging personality (i.e., physiognomic beliefs) are widespread. Here, we test how these beliefs vary across personality dimensions. Are some traits believed to be more reflected in facial features than others? Trait impressions are, to a large extent, based on the resemblance of facial features to emotion expressions. As emotional expressiveness is a central component of sociability, we hypothesized that people would more readily perceive sociability in faces. Across three preregistered studies, we find that facial features are believed to be more indicative of a person's sociability than their morality or competence. Moreover, this has consequences for the influence of facial cues in judgment and decision-making. People are more confident in the accuracy of their trait impressions when judging sociability (vs. morality or competence; Study 5.1,  $n = 527$ ), they value information on the facial appearance of job candidates more when looking for a sociable (vs. moral or competent) employee (Study 2,  $n = 390$ ), and they view reliance on facial appearance when making hiring decisions as more appropriate and more effective when looking for a sociable (vs. moral or competent) employee (Study 3,  $n = 519$ ). Together, our results provide converging evidence that people view facial appearance as especially relevant for judging a person's sociability.

People spontaneously infer personality traits from facial features (Klapper et al., 2016; Stewart et al., 2012). Even though these inferences are generally inaccurate (Bonnefon et al., 2017; Todorov, Olivola, et al., 2015), lay beliefs in the diagnostic value of facial appearance for inferring a person's character are widespread (Jaeger, Evans, Stel, & van Beest, 2019b; Suzuki et al., 2017). That is, many people believe in the core tenet of physiognomy (cf. Aristotle, 1936; Lavater, 1775)—that personality can be inferred from their facial appearance. Moreover, people rely on trait impressions when making a wide range of consequential decisions, including legal sentencing, voting behavior, and personnel selection (Olivola, Funk, et al., 2014). The effects of trait impressions on decision-making are not only prevalent, but also surprisingly persistent: People still rely on facial appearance when more diagnostic cues are available (Jaeger, Evans, Stel, et al., 2019a; Olivola et al., 2018) and when they are told to discount a person's appearance (Blair et al., 2004; Jaeger, Todorov, Evans, & van Beest, 2019).

What explains this overreliance on facial cues? Previous studies have addressed this question by examining characteristics of the perceiver (Jaeger, Evans, Stel, et al., 2019b; Suzuki, 2016; Suzuki et al., 2017). For example, Jaeger and colleagues (2019b) found that belief in physiognomy is stronger among people with an intuitive thinking style. However, physiognomic beliefs (and reliance on facial cues in general) may also vary across different personality traits. In other words, people may believe that some traits are more reflected in facial features than others. Crucially, if facial features are seen as more diagnostic for judging some personality traits, then this could influence when people rely on facial appearance. People might rely more on trait impressions (as opposed to other information) when judging a trait that is believed to be more visible in a person's facial features.

In the present studies, we examine the perceived diagnostic value of facial appearance for inferring different personality traits. Dominant

models in person perception long held that people primarily judge others along two dimensions: warmth (representing an evaluation of a person's intentions) and competence (representing an evaluation of a person's abilities; Abele & Wojciszke, 2007; Fiske, Cuddy, Glick, & Xu, 2002). However, recent studies have shown that the warmth dimensions reflects two separable components (Goodwin et al., 2014; Landy et al., 2016; Leach et al., 2007). The sociability component represents judgments of a person's ability to form interpersonal connections and encompasses traits such as friendliness, extraversion, and likeability. The morality component represents judgments of a person's moral character and encompasses traits such as trustworthiness, honesty, and loyalty. Here, we therefore focus on inferences of sociability, morality, and competence in particular.<sup>22</sup>

### **The perceived diagnostic value of facial appearance**

Are some personality traits believed to be more reflected in facial features than others? Addressing this question requires an understanding of how trait impressions from faces are formed. Theoretical accounts of social perception highlight the central role of emotion perception (Todorov, 2017; Todorov et al., 2015; Zebrowitz, 2012, 2017). Certain facial expressions, such as smiling, are not only associated with emotional states, but also with certain traits (Caulfield, Ewing, Bank, & Rhodes, 2016; Knutson, 1996; Marsh, 2005; Sutherland, Young, & Rhodes, 2016). These trait inferences are even triggered by emotionally neutral faces that merely *resemble* an emotion expression (Adams et al., 2012; Said et al., 2009). For example, slightly upturned corners of the mouth or raised eyebrows, which can both occur due to natural variation in facial appearance, can trigger ascriptions of personality traits. Thus, perception of emotion expressions are an

---

<sup>22</sup> The three dimensions are sometimes also referred to as warmth, morality, and ability (Goodwin, 2015; Goodwin et al., 2014).

important determinant of personality judgements from emotionally neutral faces.

If trait impressions from faces are largely based on resemblances to emotion expressions, then people might more readily perceive personality traits that are more strongly associated with emotional expressiveness. Functionalist accounts of emotion expressions highlight that the primary function of emotion expressions is to coordinate social interactions (Crivelli & Fridlund, 2018; Shariff & Tracy, 2011; Van Kleef, 2010). From this perspective, emotion expressions are a tool for navigating social relationships—a skill which forms the basis for evaluations of an individual's sociability (Landy et al., 2016). Thus, emotional expressiveness may be seen as particularly indicative of a person's sociability (vs. morality or competence). Evidence from several studies support this idea. Goodwin and colleagues (2014) surveyed a wide range of trait adjectives to test which traits best distinguish between judgments on the three dimensions. Dispositional happiness emerged as a defining feature of sociability judgments. In a similar vein, smiling (as opposed to displaying a neutral facial expression) has a positive impact on a wide range of trait judgments, but effects tend to be strongest for sociability judgments (Krumhuber et al., 2007; Mehu, Little, & Dunbar, 2007).

If perceptions of emotion expressions are (a) a key determinant of trait impressions from faces and (b) especially relevant for judging an individual's sociability, then people should more readily perceive sociability (than morality or competence) in faces. Results of previous studies provide preliminary evidence in favor of this hypothesis. Jaeger and colleagues (2019b) measured lay beliefs in the diagnostic value of facial appearance for inferring for sociability, morality, and competence. They found that physiognomic beliefs were strongest for sociability. In other words, people think that sociability is more reflected in facial features than morality or competence. This salience of sociability is also

emerges in unconstrained descriptions of faces. Following attractiveness, sociability and happiness were the most frequently mentioned traits when participants could freely describe faces (Oosterhof & Todorov, 2008; see also Sutherland, Liu, et al., 2017). These results suggest that facial cues may be seen as particularly relevant for judging a person's sociability.

### **The current studies**

Here, we examine whether the perceived diagnostic value of facial appearance varies as a function of which personality trait people are judging. Specifically, we hypothesize that facial features are seen as particularly relevant for judging a person's sociability (vs. morality or competence). We also examine potential consequences of this belief. We hypothesize that facial features exert a stronger influence on judgments and decisions in situations in which evaluating a person's sociability (vs. morality or competence) is central to the perceiver's goal. In short, we predict that facial cues are more influential when a target's sociability is relevant. We test these hypotheses in three preregistered studies ( $N = 1,436$ ).

First, we examine lay beliefs in the diagnostic value of facial appearance. In all three studies, we measure physiognomic beliefs for sociability, morality, and competence. In line with previous studies (Jaeger, Evans, Stel, et al., 2019b), we predict that sociability is believed to be more reflected in facial features than morality or competence. Second, we examine confidence in the accuracy of trait impressions based on facial photographs (Study 5.1,  $n = 527$ ). We measure impressions of sociability, morality, and competence and ask participants to indicate how accurate they think their judgments are. We predict that people are more confident in the accuracy of their sociability judgments (vs. their morality or competence judgments).



Third, we examine the perceived diagnostic value of facial appearance in a more applied setting. We focus on a personnel selection context, as previous studies have shown that hiring decisions are influenced by the facial appearance of candidates (Bóo et al., 2013; Gomulya et al., 2017; Ling et al., 2019). In Study 5.2 ( $n = 390$ ), we test whether the perceived diagnostic value of facial appearance for making hiring decisions varies as a function of the desired personality of job candidates. We predict that facial photographs are valued more when looking for a sociable (vs. moral or competent) employee. In Study 5.3 ( $n = 519$ ), we test whether people evaluate human resources managers who rely on facial appearance to make hiring decisions differently, depending on which personality trait the manager is looking for in candidates. Again, we predict that reliance on facial appearance is seen as more appropriate and more effective when looking for a sociable (vs. moral or competent) employee.

Next to examining differences across personality dimensions, we also investigate whether individual differences in physiognomic beliefs are related to how people form and think about trait impressions from faces. We test whether people who more strongly endorse physiognomic beliefs (across the three personality dimensions) are more confident in the accuracy of their trait impressions (Study 5.1), perceive photos as more useful for making hiring decisions (Study 5.2), and perceive reliance on photos as more appropriate and more effective when making hiring decisions (Study 5.3).

All data, analysis scripts, materials, and preregistration documents are available at the Open Science Framework (<https://osf.io/cbsmw/>). We report how our sample sizes were determined, and mention all data exclusions and measures for each study.

## Study 5.1: Confidence in judgments

In Study 5.1, we measured physiognomic beliefs for different personality traits. In line with previous studies (Jaeger, Evans, Stel, et al., 2019b), we predicted that sociability is believed to be more reflected in facial features than morality or competence. We also asked participants to judge others based on facial photographs and measured their confidence in the accuracy of their impressions. We measured perceptions of sociability, morality, and competence and predicted that participants would be more confident in their sociability judgments than their competence or morality judgments. Previous work has shown that confidence in personality inferences from faces is related to various judgment characteristics, such as judgment extremity and speed (Ames et al., 2010; Willis & Todorov, 2006). We therefore measured the extremity and speed of judgments, as these factors may also differ across personality dimensions.

In addition to examining differences across personality dimensions, we also tested whether individual differences in physiognomic beliefs are related to increased confidence. We predicted that people who more strongly endorse physiognomic beliefs are more confident in the accuracy of their judgments. In short, we examined whether confidence in judgment accuracy depended on dispositional factors (who is making the judgment?) and situational factors (which trait is being judged?).

### Methods

**Participants.** An a priori power analysis showed that a sample size of 193 participants is required to detect a small-to-medium-sized correlation between physiognomic beliefs and confidence in judgments ( $r = .200$ ) with 80% power and an alpha of 5%. We therefore aimed to recruit at least 193 participants, with the final sample size being determined by how many students participated in the study in two

weeks. A total of 533 first-year psychology students from a Dutch university completed the study in return for partial course credit. Data from five participants who always indicated the same rating across all trials were excluded from analysis, leaving a final sample of 527 participants ( $M_{age} = 19.60$ ,  $SD_{age} = 2.09$ ; 81.21% female, 18.60% male, 0.19% other).<sup>23</sup> The majority of participants were Dutch (73.62%) or German (13.28%).

**Materials and procedure.** We selected ten images from the Chicago Face Database (five male, five female) and asked participants to rate the images on three dimensions: *sociability* (*warmth, friendliness, likeability*), *morality* (*trustworthiness, sincerity, honesty*), and *competence* (*competence, intelligence, skillfulness*). Participants rated each image on each dimension in a random order on a scale from 1 (*not at all* [dimension]) to 9 (*extremely* [dimension]). Next to recording the strength of ratings (i.e., how high or low they scored each face in a given dimension), we also recorded their speed and extremity (i.e., the absolute distance to the midpoint of our scale). After each judgment, participants indicated their confidence in the accuracy of their judgment on a scale from 1 (*not at all confident*) to 9 (*extremely confident*).

Participants then filled out several unrelated questionnaires and, approximately ten minutes later, completed the physiognomic beliefs scale (Jaeger, Evans, Stel, et al., 2019b). The questions were preceded by a statement that prompted participants to “imagine seeing the passport photo of a stranger”. We measured general belief with three questions (e.g., “I can learn something about a person’s personality just from looking at his or her face”; Cronbach’s  $\alpha = .90$ ). We also measured belief for three specific trait dimensions (sociability, competence, and morality) with three traits for each dimension (sociability: warmth, likeability, and friendliness,  $\alpha = .90$ ; competence: competence,

---

<sup>23</sup> We did not specify missing data as an exclusion criteria in our preregistration. Retaining participants with missing data did not influence the pattern of results.

intelligence, and skillfulness  $\alpha = .90$ ; morality: trustworthiness, honesty, and sincerity  $\alpha = .89$ ; Brambilla et al., 2011). Participants dragged a slider to indicate how accurately they think they can judge each trait just from looking at a person's face. Slider values range from 0 (*not accurately at all*) to 100 (*extremely accurate*). We included three additional characteristics not related to a person's personality (gender, age, and attractiveness).

In line with our preregistered exclusion criteria, we excluded 305 response times (1.91%) that were three standard deviations below or above the mean. Response times were  $\log_{10}$ -transformed due to their right-skewed distribution. Dutch participants completed the study in Dutch while non-Dutch participants completed the study in English.

## Results

**Descriptive statistics.** On average, participants took 6.31 seconds ( $SD = 2.50$ ) to make a judgment and the mean confidence rating was 5.27 ( $SD = 1.65$ , on a scale that ranged from 1 to 9). We computed intraclass correlation coefficients (*ICCs*) to estimate consensus in judgments across participants (Shrout & Fleiss, 1979). Participants showed significant consensus in their judgments of sociability,  $ICC(2, 1) = .394$ ,  $p < .001$ , 95% CI [.234, .684], morality,  $ICC(2, 1) = .231$ ,  $p < .001$ , 95% CI [.123, .501], and competence,  $ICC(2, 1) = .380$ ,  $p < .001$ , 95% CI [.224, .672].

**Physiognomic beliefs.** The average score on general physiognomic beliefs was just above the midpoint of our scale ( $M = 4.05$ ,  $SD = 1.26$ ),  $t(526) = 0.88$ ,  $p = .38$ ,  $d = 0.04$  (see Table 4.1). Around half of all participants (51.04%) believed at least somewhat in physiognomy (i.e., scored above the midpoint of the scale),  $\chi^2(1) = 0.19$ ,  $p = .66$ .

We compared physiognomic beliefs across the three personality dimensions to test whether people think sociability is more reflected in facial features than morality or competence. Physiognomic beliefs were significantly higher for sociability ( $M = 44.64$ ,  $SD = 22.99$ ) compared to

morality ( $M = 28.82, SD = 21.35$ ),  $t(526) = 23.02, p < .001, d = 1.00$ , and competence ( $M = 30.05, SD = 20.17$ ),  $t(526) = 19.29, p < .001, d = 0.84$  (see Figure 5.1). Morality-specific beliefs were slightly higher than competence-specific beliefs, but this difference was only marginally significant,  $t(526) = 1.78, p = .075, d = 0.20$ .

Table 5.1

Descriptive statistics for general physiognomic beliefs across the three studies.

Sample	$n$	$\alpha$	$M$	$SD$	% Believers
Study 5.1	527	.75	4.05	1.26	51.04
Study 5.2	390	.90	3.90	1.41	50.00
Study 5.3	519	.89	3.55	1.35	38.54

*Note.* General physiognomic beliefs were measured with three items that were rated on a scale from 1 (*strongly disagree*) to 7 (*strongly agree*). “% Believers” indicates the percentage of participants that scored above the midpoint of the scale.

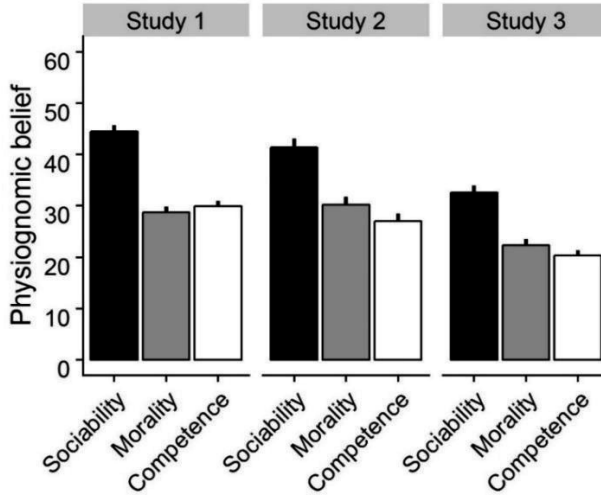


Figure 5.1. Differences in physiognomic beliefs for sociability, morality, and competence across the three studies. Error bars represent bootstrapped 95% confidence intervals.

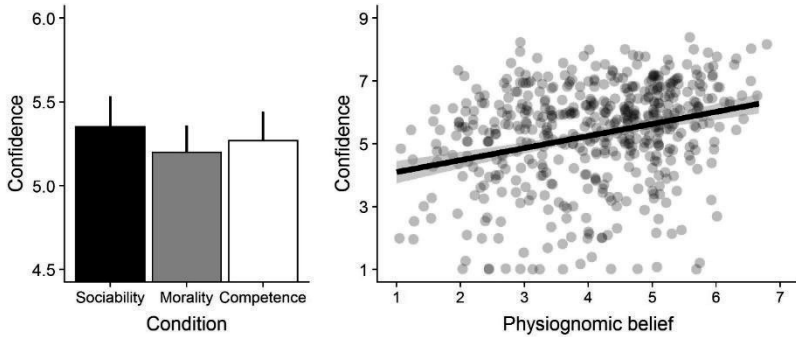
**Confidence across personality traits.** Next, we tested whether confidence levels depended on which dimension participants were judging. We predicted that participants would be more confident in their sociability judgments (vs. competence or morality judgments). To test this prediction, we estimated a multilevel regression model with random intercepts per participant and face, in which we regressed confidence on a dummy variable indicating which dimension was judged (see Figure 5.2). In line with our hypothesis, participants were more confident when judging sociability than when judging competence,  $\beta = 0.093$ ,  $SE = 0.023$ ,  $t(15,272) = 4.13$ ,  $p < .001$ , or morality,  $\beta = 0.158$ ,  $SE = 0.023$ ,  $t(15,272) = 7.02$ ,  $p < .001$ . Participants were also more confident when making competence (vs. morality) judgments,  $\beta = 0.065$ ,  $SE = 0.023$ ,  $t(15,272) = 2.89$ ,  $p = .004$ .

We also explored whether differences in confidence across the three dimensions would still emerge when controlling for the extremity

and speed of judgments. Regressing confidence on a dummy variable indicating which dimension was judged, as well as judgment extremity and speed, showed a positive effect of extremity,  $\beta = 0.383$ ,  $SE = 0.010$ ,  $t(15,069) = 39.16$ ,  $p < .001$ , and a negative effect of speed,  $\beta = -0.161$ ,  $SE = 0.010$ ,  $t(15,123) = 15.34$ ,  $p < .001$ . More importantly, there was still a significant effect of personality dimension: Participants were more confident when judging sociability than when judging competence,  $\beta = 0.079$ ,  $SE = 0.021$ ,  $t(14,967) = 3.70$ ,  $p < .001$ , or morality,  $\beta = 0.094$ ,  $SE = 0.021$ ,  $t(14,968) = 4.40$ ,  $p < .001$ . There was no difference in confidence between competence and morality judgments,  $\beta = -0.005$ ,  $SE = 0.021$ ,  $t(15,272) = 0.22$ ,  $p = .83$ .

**Physiognomic beliefs and confidence.** To test whether people who more strongly endorse physiognomic beliefs are more confident in their judgments, we correlated average confidence across all trials with physiognomic belief scores. In line with our hypothesis, there was a positive correlation between physiognomic beliefs and confidence,  $r(525) = .292$ ,  $p < .001$  (see Figure 5.2). People who more strongly endorsed physiognomic beliefs were more confident in the accuracy of their face-based personality judgments.

We also explored whether physiognomic beliefs were still related to confidence when controlling for the extremity and speed of judgments. Regressing confidence on all three predictors showed a positive effect of extremity,  $\beta = 0.478$ ,  $SD = 0.067$ ,  $t(523) = 7.16$ ,  $p < .001$ , and no effect of speed,  $\beta = -0.034$ ,  $SD = 0.066$ ,  $t(523) = 0.51$ ,  $p = .61$ . Participants were more confident when they made higher and more extreme judgments, but not when making faster judgments. Crucially, there was a positive effect of physiognomic beliefs,  $\beta = 0.409$ ,  $SD = 0.067$ ,  $t(523) = 6.12$ ,  $p < .001$ , showing that physiognomic beliefs still predicted confidence when controlling for the extremity and speed of judgments.



*Figure 5.2.* Differences in judgment confidence between judgments of sociability, competence, and morality (left). Correlations between physiognomic beliefs and judgment confidence (right). Error bars and ribbons represent bootstrapped 95% confidence intervals.

## Discussion

Replicating previous findings (Jaeger, Evans, Stel, et al., 2019b), we found that sociability is believed to be more reflected in facial features than morality or competence. Results of the current study showed that this pattern extends to how confident people are in the accuracy of their trait impressions from faces. We found that participants were more confident in the accuracy of their sociability impressions than their morality or competence impressions. Differences in confidence across the three dimensions still emerged when controlling for the extremity, speed, and strength of judgments, which were all related to confidence.

### Study 5.2: Diagnostic value of facial appearance

Results of Study 5.1 showed that sociability is not only believed to be more reflected in facial features than morality or competence, participants were also more confident in the accuracy of their sociability judgments. Going beyond impressions formation based on facial photographs, Study 5.2 examined the perceived diagnostic value of facial appearance in a more applied setting. Previous studies have shown that



hiring decisions are influenced by the facial appearance of candidates (Bóo et al., 2013; Gomulya et al., 2017; Ling et al., 2019). In the current study, we tested if the diagnostic value of facial appearance in a personnel selection context depends on which personality trait people are looking for in candidates. As sociability is believed to be more reflected in faces than morality or competence, we predicted that people would value information on a candidate's facial appearance more if they are looking for sociable (vs. moral or competent) employee.

We also measured physiognomic beliefs to test if the finding that sociability is believed to be more reflected in facial features than morality or competence replicates in a sample of participants from the United States. This also allowed us to test whether physiognomic beliefs are related to the perceived value of personal photos for making hiring decisions. Specifically, we predicted that people who more strongly believe that personality is reflected in

## Methods

**Participants.** An a priori power analysis showed that a sample size of 130 participants per condition is required to detect a small-to-medium-sized difference between condition ( $d = 0.35$ ) with 80% power (and  $\alpha = 5\%$ ). We therefore aimed to recruit a total of 390 participants. In total, we recruited 430 US American MTurk workers who completed the study in return for \$0.50 each. In line with our preregistration, data from 2 participants (0.47%) who indicate poor or basic English proficiency and from 38 participants (8.84%) who failed an attention check at the end of the study were excluded from analysis, leaving a final sample of 390 participants ( $M_{age} = 34.64$ ,  $SD_{age} = 9.88$ ; 42.56% female).

**Materials and procedure.** We asked participants to imagine that they are working in the HR department of a company that hosts various events and they were tasked with hiring a new event planner. Participants were randomly assigned to one of three conditions, which

determined which personality trait they were looking for in candidates (sociability, morality, or competence). For instance, in the sociability condition, participants read: “In the past, your company has received complaints about planners who were unfriendly and dismissive to guests. Therefore, you are looking for someone who is warm, friendly, and likeable”.

To measure the perceived value of a facial photograph, participants saw a list of five CV components (photograph, experience, recommendation letter, education level, grades) in a random order and were asked to rate how useful each component is for making the hiring decision on a scale from 0 (*not at all useful*) to 100 (*extremely useful*). The usefulness ratings of the applicant’s photograph constitutes our key dependent variable. Finally, participants completed the physiognomic beliefs scale. As in Study 5.1, we measured general physiognomic belief ( $\alpha = .90$ ) and specific physiognomic beliefs for sociability ( $\alpha = .93$ ), morality ( $\alpha = .95$ ), and competence ( $\alpha = .93$ ).

## Results

**Descriptive statistics.** Usefulness ratings were lower for photos ( $M = 33.15, SD = 28.33$ ) than for recommendation letters ( $M = 80.06, SD = 17.77$ ), past experience ( $M = 66.63, SD = 26.49$ ), level of education ( $M = 50.91, SD = 28.27$ ), and grades ( $M = 43.15, SD = 27.73$ ), all  $p < .001, 0.35 < d < 1.98$ .

**Physiognomic beliefs.** The average score on general physiognomic beliefs was just below the midpoint of our scale ( $M = 3.90, SD = 1.41$ ),  $t(389) = 1.39, p = .17, d = 0.07$ . Half of all participants (50.00%) believed at least somewhat in physiognomy (i.e., they scored above the midpoint of the scale).

We compared physiognomic beliefs across the three personality dimensions to test whether people think sociability is more reflected in facial features than morality or competence (see Figure 5.1). Again,

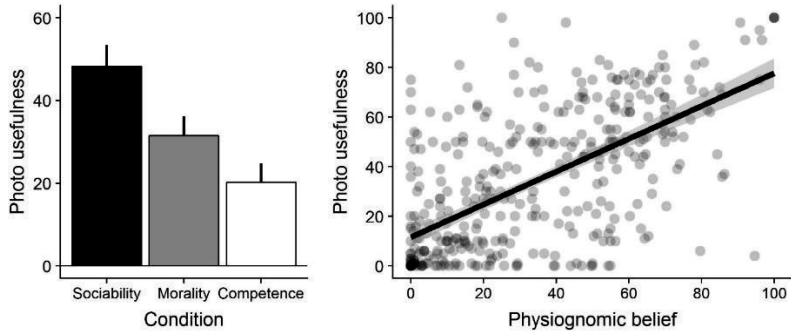
physiognomic beliefs were significantly higher for sociability ( $M = 41.42$ ,  $SD = 27.67$ ) compared to morality ( $M = 30.21$ ,  $SD = 26.48$ ),  $t(389) = 15.42$ ,  $p < .001$ ,  $d = 0.78$ , and competence ( $M = 26.99$ ,  $SD = 24.59$ ),  $t(389) = 16.47$ ,  $p < .001$ ,  $d = 0.83$  (see Figure 5.1). Morality-specific beliefs were slightly higher than competence-specific beliefs, but this difference was less pronounced,  $t(389) = 5.72$ ,  $p < .001$ ,  $d = 0.29$ .

**Photo usefulness across personality traits.** Next, we tested whether usefulness ratings of the photo depended on which personality trait participants were looking for in candidates. As participants believed that sociability is more reflected in facial features than morality or competence, we predicted that they would rate photos as more important for more making hiring decisions when they were looking for a sociable (vs. moral or competent) candidate. Usefulness ratings of photos varied across the three conditions  $F(2, 378) = 37.42$ ,  $p < .001$  (see Figure 5.3). In line with our hypothesis, participants perceived photos as a more useful cue when looking for a sociable candidate ( $M = 48.27$ ,  $SD = 27.42$ ) than when looking for a moral candidate ( $M = 31.53$ ,  $SD = 26.55$ ),  $t(245.2) = 8.69$ ,  $p < .001$ ,  $d = 1.09$ , or a competent candidate ( $M = 20.19$ ,  $SD = 23.93$ ),  $t(255.5) = 5.00$ ,  $p < .001$ ,  $d = 0.62$ . Participants perceived photos as a more useful cue when looking for a moral (vs. competent) candidate,  $t(245.2) = 8.69$ ,  $p < .001$ ,  $d = 0.45$ . Thus, people not only believed that sociability is more reflected in faces than morality or competence, they also saw photos as more important information for making hiring decisions when looking for a sociable (vs. moral or competent) candidate.

**Photo usefulness and physiognomic beliefs.** We also explored whether individual differences in physiognomic beliefs were associated with the perceived usefulness of photos. We correlated usefulness ratings with the conditions-specific physiognomic belief score. For example, for participant in the sociability condition, we correlated usefulness ratings with sociability-specific physiognomic beliefs. Across

the three conditions, this revealed a positive correlation between physiognomic beliefs and usefulness ratings for photos,  $r(388) = .631$ ,  $p < .001$  (see Figure 5.3). In other words, people who scored high on physiognomic beliefs saw personal photos as more useful for making hiring decisions.

Finally, we explored associations between physiognomic beliefs and usefulness ratings of the other cues. Physiognomic beliefs reflect the perceived diagnostic value of a person's facial appearance. We would therefore predict that they are related to the perceived usefulness of facial photographs, but not to other cues such as a recommendation letter. To test this prediction, we regressed physiognomic beliefs on usefulness ratings for all five cues (all variables were  $z$ -standardized prior to analysis). Physiognomic beliefs were positively associated with usefulness ratings for photos,  $\beta = 0.613$ ,  $SD = 0.039$ ,  $t(384) = 15.69$ ,  $p < .001$ . Surprisingly, they were also positively associated with usefulness ratings for experience,  $\beta = 0.136$ ,  $SD = 0.067$ ,  $t(522) = 2.27$ ,  $p = .024$ , and negatively associated with usefulness ratings for level of education,  $\beta = -0.179$ ,  $SD = 0.048$ ,  $t(384) = 3.74$ ,  $p < .001$ . There were no significant relationships between physiognomic beliefs and usefulness ratings for recommendation letters,  $\beta = -0.008$ ,  $SD = 0.040$ ,  $t(384) = 0.21$ ,  $p = .84$ , and grades,  $\beta = 0.069$ ,  $SD = 0.055$ ,  $t(384) = 1.25$ ,  $p = .21$ . Importantly, physiognomic beliefs were more strongly associated with usefulness ratings of photos than with usefulness ratings of the other cues (all  $z > 6.68$ , all  $p < .001$ ).



*Figure 5.3.* Differences in rated usefulness of personal photos for making hiring decisions as a function of desired trait in the candidate (left). Correlation between physiognomic beliefs and rated usefulness of personal photos for making hiring decisions (right). Error bars and ribbons represent bootstrapped 95% confidence intervals.

## Discussion

We again replicated the finding that sociability is believed to be more reflected in facial features than morality or competence. Crucially, this pattern was also reflected in how much participants valued information on a candidate’s facial appearance when making hiring decisions. Personal photos were seen as more useful when participants were looking for a sociable (vs. moral or competent) employee. We also found that individual differences in physiognomic beliefs were related to the perceived diagnostic value of photos. The stronger participants endorsed the belief that personality is reflected in facial features, the more they valued personal photos for making hiring decisions.

### Study 5.3: Appropriateness and effectiveness of relying on facial appearance

Studies 5.1 and 5.2 provided insights into how much people value facial cues when judging different personality traits. In Study 5.3, we extended our analyses to a third-party perspective. That is, we examined

how people evaluate others who rely on facial cues. We again tested this in a personnel selection context. Specifically, we tested whether attitudes towards appearance-based discrimination vary depending on which personality trait is inferred from facial appearance. As sociability is believed to be more reflected in faces than morality or competence, we predicted that people would find reliance on personal photos more acceptable and more effective when looking for a sociable (vs. moral or competent) candidate.

We again measured physiognomic beliefs and aimed to replicate the finding that sociability is believed to be more reflected in facial features than morality or competence. This also allowed us to test whether physiognomic beliefs are related to attitudes towards appearance-based discrimination. Specifically, we predicted that people who more strongly believe that personality is reflected in facial features would perceive reliance on personal photos to make hiring decisions as more appropriate and more effective.

## Methods

**Participants.** An a priori power analysis showed that a sample size of 173 participants per condition is required to detect a small-to-medium-sized difference between condition ( $d = 0.35$ ) with 90% power (and  $\alpha = 5\%$ ). We therefore aimed to recruit a total of 519 participants. In total, we recruited 621 US American MTurk workers who completed the study in return for \$0.50 each. In line with our preregistration, data from 7 participants who indicate poor or basic English proficiency and from 93 participants who failed at least one of two attention checks were excluded from analysis, leaving a final sample of 521 participants ( $M_{age} = 35.78$ ,  $SD_{age} = 10.93$ ; 45.11% female, 54.51% male, 0.38% other).

**Materials and procedure.** We asked participants to imagine a manager who is working in the HR department of a company that hosts various events. The job of the manager is to hire event planners.

Participants were randomly assigned to one of three conditions, which determined which personality trait the manager is looking for in candidates (sociability, morality, or competence). For instance, in the sociability condition, participants read:

In the past, the company has received complaints about planners who were unfriendly and dismissive to guests. Therefore, the manager is looking for someone who is warm, friendly, and likeable. Since the manager is receiving many applications, he always selects a few promising candidates that are then invited for an interview. When deciding whom to invite, he looks at the candidates' photos and picks the ones that look very warm, friendly, and likeable to him.

To measure attitudes towards discrimination based on facial photographs, we asked participants to evaluate the appropriateness and effectiveness of the manager's practice. Participants rated the appropriateness of the strategy with three items (appropriate, moral, ethical;  $\alpha = 0.95$ ) on a scale from -50 (e.g., *extremely inappropriate*) to 50 (e.g., *extremely appropriate*). Participants rated the effectiveness of the strategy with three items (effective, successful, likely to achieve goal;  $\alpha = 0.97$ ) on a scale from -50 (e.g., *extremely ineffective*) to 50 (e.g., *extremely effective*). The order of appropriateness and effectiveness ratings was randomized. Finally, participants completed the physiognomic beliefs scale. As in Study 5.1, we measured general physiognomic belief ( $\alpha = .89$ ) and specific physiognomic beliefs for sociability ( $\alpha = .94$ ), morality ( $\alpha = .94$ ), and competence ( $\alpha = .94$ ).

## Results

**Descriptive statistics.** On average, the manager's strategy to make hiring decisions based on photos was seen as relatively inappropriate ( $M = -15.54$ ,  $SD = 26.24$ ),  $t(520) = 13.52$ ,  $p < .001$ ,  $d = 0.59$ , and ineffective ( $M = -14.36$ ,  $SD = 25.18$ ),  $t(520) = 13.02$ ,  $p < .001$ ,  $d = 0.57$ .

Appropriateness ratings and effectiveness ratings were strongly correlated,  $r(519) = .738, p < .001$ .

**Physiognomic beliefs.** The average score on general physiognomic beliefs was just below the midpoint of our scale ( $M = 3.55, SD = 1.35$ ),  $t(520) = 7.63, p < .001, d = 0.33$ . Around one-third of all participants (38.58%) believed at least somewhat in physiognomy (i.e., they scored above the midpoint of the scale),  $\chi^2(1) = 26.73, p < .001$ .

We compared physiognomic beliefs across the three personality dimensions to test whether people think sociability is more reflected in facial features than morality or competence (see Figure 5.1). Again, physiognomic beliefs were significantly higher for sociability ( $M = 32.55, SD = 26.16$ ) compared to morality ( $M = 22.23, SD = 22.77$ ),  $t(520) = 16.67, p < .001, d = 0.73$ , and competence ( $M = 20.21, SD = 21.74$ ),  $t(520) = 17.45, p < .001, d = 0.76$  (see Figure 5.1). Morality-specific beliefs were slightly higher than competence-specific beliefs, but this difference was less pronounced,  $t(520) = 4.56, p < .001, d = 0.20$ .

**Appropriateness and effectiveness ratings across personality traits.** Next, we tested whether appropriateness and effectiveness ratings of the manager's strategy depended on which personality trait the manager was looking for in candidates. We predicted that participants would rate relying on photos to make hiring decisions as more appropriate and effective when the manager is looking for a sociable (vs. moral or competent) candidate. Appropriateness ratings varied across the three conditions  $F(2, 518) = 11.00, p < .001$  (see Figure 5.4). In line with our hypothesis, participants perceived relying on photos as a more appropriate strategy when looking for a sociable candidate ( $M = -9.77, SD = 26.24$ ) than when looking for a competent candidate ( $M = -22.48, SD = 24.51$ ),  $t(340.1) = 4.66, p < .001, d = 0.50$ . However, the difference in appropriateness ratings between the sociability and morality condition ( $M = -14.04, SD = 26.49$ ) was not significant,  $t(340.9) = 1.50, p = .14, d = 0.16$ . Participants perceived



relying on photos as a more appropriate strategy when looking for a moral (vs. competent) candidate,  $t(346.5) = 3.10, p = .002, d = 0.33$ .

Effectiveness ratings also varied across the three conditions  $F(2, 518) = 16.35, p < .001$  (see Figure 5.5). In line with our hypothesis, participants perceived relying on photos as a more effective strategy when looking for a sociable candidate ( $M = -5.77, SD = 25.06$ ) than when looking for a moral candidate ( $M = -16.67, SD = 23.56$ ),  $t(338.2) = 4.15, p < .001, d = 0.45$ , or a competent candidate ( $M = -20.26, SD = 24.76$ ),  $t(343.6) = 5.41, p < .001, d = 0.58$ . There was no significant difference in effectiveness ratings between the morality and competence condition,  $t(349.7) = 1.39, p = .16, d = 0.15$ .

Thus, people not only believed that sociability is more reflected in faces than morality or competence, they also saw reliance on photos as more appropriate and effective for making hiring decisions when looking for a sociable (vs. moral or competent) candidate.

**Appropriateness and effectiveness ratings and physiognomic beliefs.** We also tested whether individual differences in physiognomic beliefs were associated with the perceived appropriateness and effectiveness of the manager's strategy. We correlated appropriateness and effectiveness ratings with the condition-specific physiognomic belief score. For example, for participant in the sociability condition, we correlated appropriateness and effectiveness ratings with sociability-specific physiognomic beliefs. This revealed a positive correlation between physiognomic beliefs and appropriateness ratings,  $r(519) = .530, p < .001$  (see Figure 5.4), and effectiveness ratings,  $r(519) = .659, p < .001$  (see Figure 5.5). In other words, people who scored high on physiognomic beliefs found it more appropriate and more effective to rely on personal photos for making hiring decisions.

Appropriateness and effectiveness ratings were strongly correlated. Therefore, we also explored whether they shared unique variance with physiognomic beliefs. We found a significant relationship

between physiognomic beliefs and the perceived effectiveness of relying on photos when controlling for perceived appropriateness,  $\beta = 0.380$ ,  $SE = 0.032$ ,  $t(518) = 12.04$ ,  $p < .001$ . In a similar vein, we found a significant relationship between physiognomic beliefs and the perceived appropriateness of relying on photos when controlling for perceived effectiveness,  $\beta = 0.082$ ,  $SE = 0.042$ ,  $t(518) = 1.97$ ,  $p = .049$ .

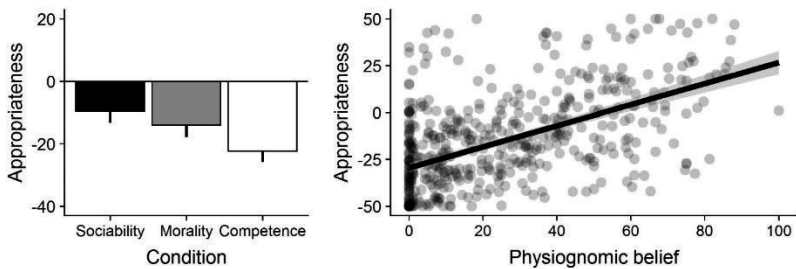


Figure 5.4. Differences in rated appropriateness of relying on photos to make hiring decisions as a function of desired trait in the candidate (left) and correlation between physiognomic beliefs and appropriateness ratings across conditions (right). Error bars and ribbons represent bootstrapped 95% confidence intervals.

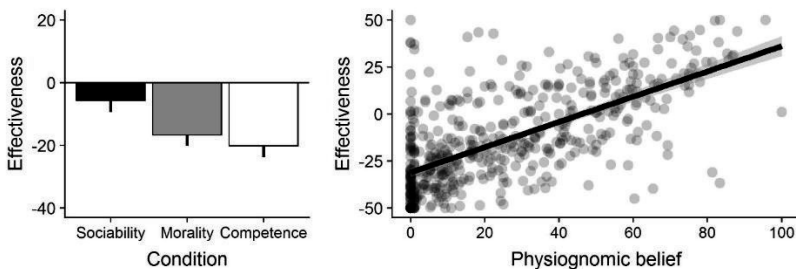


Figure 5.5. Differences in rated effectiveness of relying on photos to make hiring decisions as a function of desired trait in the candidate (left) and correlation between physiognomic beliefs and effectiveness ratings across conditions (right). Error bars and ribbons represent bootstrapped 95% confidence intervals.

## **Discussion**

We again replicated the finding that sociability is believed to be more reflected in facial features than morality or competence. Crucially, this pattern was also reflected in how participants evaluated the appropriateness and effectiveness of relying on facial appearance for making hiring decisions. Even though participants perceived this strategy as less appropriate and less effective than reliance on other information (e.g., recommendation letters), attitudes differed depending on which personality trait recruiters were looking for in candidates. Specifically, reliance on personal photos for making hiring decisions was seen as more appropriate and more effective when looking for a sociable (vs. moral or competent) employee.

We also found that individual differences in physiognomic beliefs were related to attitudes towards appearance-based discrimination. Participants who more strongly endorsed the belief that personality is reflected in facial features, perceived reliance on facial appearance for making hiring decisions as more appropriate and more effective.

## **General discussion**

Many people believe in the central idea of physiognomy that a person's facial appearance is indicative of their personality (Jaeger, Evans, Stel, et al., 2019b; Suzuki et al., 2017). Here, we asked whether some personality traits are believed to be more reflected in facial features than others. Trait impressions are, to a large extent, based on the resemblance of facial features to emotion expressions (Adams et al., 2012; Said et al., 2009). As emotional expressiveness is a central component of sociability (Kring et al., 1994; Roger & Neshs Hoover, 1987), we hypothesized that people would more readily perceive sociability in faces than morality or competence. In line with this hypothesis, we found that sociability is believed to be more reflected in facial features than morality or competence. This effect replicated across three large

samples of Dutch (Study 5.1) and U.S. American participants (Studies 5.2 and 5.3) and is in line with previous findings (Jaeger, Evans, Stel, et al., 2019b).

Going beyond previous investigations, we examined the consequences of this belief. In particular, we investigated whether facial cues are more influential when a person's sociability (vs. morality or competence) is relevant. Across three studies, we manipulated which trait was relevant for participants' judgments or decisions. We measured the perceived diagnostic value of facial appearance by assessing (a) how confident participants are in the accuracy of their face-based personality judgments (Study 5.1), (b) how much participants value information on the facial appearance of job candidates when making hiring decisions (Study 5.2), and (c) how appropriate and effective participants view reliance on facial appearance when making hiring decisions (Study 5.3).

Study 5.1 showed that participants were more confident in the accuracy of their sociability judgments (vs. morality or competence judgments). In Study 5.2, we examined how much people value information on a person's facial appearance (i.e., a facial photograph) in a hiring context. While photographs of candidates were seen as less diagnostic than other cues (e.g., recommendation letters or past experience), photographs were valued more when looking for a sociable (vs. moral or competent) employee. A similar pattern of results was found in Study 5.3. Participants viewed a manager's strategy to rely on facial appearance when making hiring decisions as relatively inappropriate and ineffective. However, reliance on facial appearance was seen as more appropriate and more effective when looking for a sociable (vs. moral or competent) employee. Together, we find converging evidence for the notion that facial appearance is seen as particularly important for judging a person's sociability.

## Theoretical implications

People often rely on facial appearance even when more diagnostic information is available (Jaeger, Evans, Stel, et al., 2019a; Olivola et al., 2018). Overreliance on facial cues is problematic because it can lead to worse outcomes for the decision-maker (Olivola & Todorov, 2010b) and to systematic biases against people with a certain facial appearance (Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006; J. P. Wilson & Rule, 2015). To explain the widespread influence of facial appearance, previous studies have mostly focused on individual characteristics (Ewing, Caulfield, Read, & Rhodes, 2015; Jaeger, Evans, Stel, et al., 2019b; Suzuki, 2016). The present findings suggest that the extent to which facial appearance influences decision-making outcomes not only depends on who is making the decision, but also in what context the decision is being made. People may be more likely to rely on trait impressions from faces when a person's sociability (vs. morality or competence) is relevant for their decision. This implies that decision-makers need to be aware of the biasing influence of facial appearance, especially when judging a person's sociability.

Our results also extend previous findings on how individual characteristics influence reliance on facial cues. Jaeger and colleagues (2019b) found that people who believe in physiognomy are more confident in the accuracy of their trait impressions and rely more on trait impressions when making trust decisions. Study 1 replicated the positive association between physiognomic beliefs and confidence across various personality traits. Moreover, Studies 5.2 and 5.3 showed that individual differences in physiognomic beliefs are associated with the weight people attach to facial cues in a personnel selection context. Participants who more strongly endorsed physiognomic beliefs (a) valued personal photos more for making hiring decisions and (b) viewed a manager's strategy to make selection decisions based on personal photos as more appropriate and more effective. Together, the current results suggest

that the extent to which facial cues affect judgments and decisions depends on dispositional factors (who is making the judgment?) and situational factors (which personality trait is being judged?).

Studies in the field of social perception often highlight the efficient and automatic way in which people form trait impressions from faces (Bonnefon et al., 2013; Engell et al., 2007; Willis & Todorov, 2006). However, these findings should not be interpreted as evidence for procedural uniformity in impression formation. In fact, recent studies have shown that there are various top-down influences on social perception (Freeman & Ambady, 2010; Hehman, Stoler, Freeman, Flake, & Xie, 2019). For instance, lay beliefs about the conceptual overlap between specific personality traits (e.g., whether trustworthiness correlates with sociability) predict the extent to which face-based impressions of the traits overlap (e.g., whether trustworthiness impressions correlate with sociability impressions; Stoler, Hehman, & Freeman, 2018; Stoler, Hehman, Keller, et al., 2018). In a similar vein, our results show that many aspects of the impression formation process (e.g., the speed and extremity of judgments) are influenced by characteristics of the perceiver and the situation.

### **Limitations and future directions**

Why is facial appearance seen as especially diagnostic for judging sociability? We suggested that this is due to the fact that trait impressions are primarily formed based on emotion cues (i.e. facial features that resemble emotion expressions). As emotional expressiveness is a central component of sociability, people more readily perceive sociability in faces. An alternative (albeit non-mutually exclusive) explanation holds that people rely more on facial appearance when judging sociability because sociability can be more accurately inferred than other personality traits. Previous research using stranger rating paradigms—in which participants make personality judgments

based on limited information—suggests that judgment accuracy is highest for extraversion (which is conceptually similar to sociability) compared to other Big Five personality traits (Kenny & West, 2008).

However, the evidence is less clear for judgments based on static images of emotionally neutral faces. Even though some studies found that extraversion judgments are somewhat accurate, evidence for the claim that extraversion judgments are more accurate than other personality trait judgments is mixed (Borkenau et al., 2009; Kramer & Ward, 2010; Naumann et al., 2009; Penton-Voak et al., 2006). Moreover, other studies found no evidence for accuracy in extraversion judgments (Ames et al., 2010; A. L. Jones et al., 2012; Shevlin et al., 2003). More research is needed to determine the accuracy of different personality trait judgments from facial features and to what extent this pattern is reflected in people's beliefs about the accuracy of their own trait judgments.

Previous studies suggest that people are generally not aware of how accurate their judgments are—that is, they show poor meta-accuracy (Ames et al., 2010; Jaeger, Evans, Stel, et al., 2019b). People who are more confident are on average not more accurate and, at the individual rating level, people's confidence does not seem to track their judgment accuracy. Nevertheless, people might show meta-accuracy at the trait level. That is, people might be more confident in their trait judgments when judging traits that can actually be inferred with higher levels of accuracy. To test this hypothesis, future studies could measure accuracy and confidence for a wide range of personality traits.

## Conclusion

People spontaneously infer personality traits from faces and rely on their inferences to make a wide range of social decisions (Olivola, Funk, et al., 2014; Todorov, Olivola, et al., 2015). Here, we asked whether the perceived diagnostic value of facial appearance varies across

different personality dimensions. Are some traits believed to be more reflected in facial features than others? We found that beliefs in the manifestation of personality traits in facial features are strongest for sociability (vs. morality or competence). This belief has consequences for the weight people attach to a person's facial appearance when judging different personality traits. We found that people are more confident in the accuracy of their trait impressions when judging sociability (vs. morality or competence). In a hiring context, people value information on a person's facial appearance more when looking for a sociable (vs. moral or competent) employee. Finally, reliance on facial appearance to select job candidates is seen as more appropriate and more effective when looking for a sociable (vs. moral or competent) employee. Overall, the current results provide converging evidence that people see facial appearance as especially relevant for judging a person's sociability. Put differently, facial cues weigh heavy in judgment and decision-making, especially when judging another person's sociability



# Chapter 6

## Explaining the persistent influence of facial cues in social decision-making

Based on:

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.

All data, preregistration documents, and analysis scripts are available at the Open Science Framework (<https://osf.io/h6dsj/>).

### **Abstract**

Impressions of trustworthiness based on facial cues influence many consequential decisions, in spite of their (generally) poor accuracy. Here, we test whether reliance on facial cues can be better explained by (a) the belief that facial cues are more valid than other cues or by (b) the quick and primary processing of faces, which makes relying on facial cues relatively effortless. Six studies ( $N = 2,732$  with 73,182 trust decisions) test the two accounts by comparing the effects of facial cues and economic payoffs on trust decisions. People believe that facial cues are less valid than economic payoffs (Study 6.1), but relying on facial cues takes less time than relying on economic payoffs (Study 6.2). Critically, introducing facial cues causes people to discount payoff information, but introducing payoff information does not reduce the effect of facial cues (Studies 6.3-6.5). Finally, when making intuitive (vs. reflective) trust decisions, people rely less on payoff information, but they do not rely less on facial cues (Study 6.6). Together, these findings suggest that persistent reliance on facial trustworthiness is better explained by the intuitive accessibility of facial cues, rather than beliefs that facial cues are particularly valid.

First impressions are often based solely on a person's facial appearance (Oosterhof & Todorov, 2008). A number of studies have demonstrated that these face judgments, especially judgments of trustworthiness, influence decisions in domains such as voting, personnel selection, and criminal sentencing (Olivola, Funk, et al., 2014; Olivola & Todorov, 2010a; Todorov, Olivola, et al., 2015). The widespread influence of facial trustworthiness is surprising, given that the human ability to accurately identify trustworthy individuals based on facial cues is generally poor (Todorov, Olivola, et al., 2015).

Why do people rely on first impressions based on facial cues? On the one hand, models of cue selection propose that the perceived diagnosticity of a cue (i.e., how well it is thought to predict a certain outcome) determines how much people rely on it when making decisions (Brunswik, 1956; Gigerenzer & Goldstein, 1996; Hammond, Hirsch, & Todd, 1964). Consequently, the influence of facial cues has been attributed to the *perceived* validity of face judgments (Olivola, Funk, et al., 2014; Rezlescu et al., 2012): This view suggests that people rely on facial trustworthiness judgments because they *believe* that their judgments are accurate. On the other hand, it has also been suggested that face judgments affect decisions because of their intuitive accessibility (Olivola & Todorov, 2010a; Willis & Todorov, 2006). Faces are processed quickly and efficiently (Freeman & Johnson, 2016) and people tend to prioritize cues that come to mind easily (Evans & Krueger, 2016; Shah, 2007; Simmons & Nelson, 2006). Relying on facial cues might therefore constitute a mental shortcut that reduces decision effort (Gigerenzer, Hertwig, & Pachur, 2011; Shah & Oppenheimer, 2008). Here, we test these two accounts, illuminating the mechanisms that give rise to the widespread influence of facial cues.

### Facial trustworthiness influences social decision-making

The ability to judge trustworthiness in others is crucial in mixed-motive settings, where there is a motivational conflict between self-interest and the collective good (Kelley et al., 2003). When this conflict arises, people must judge whether their interaction partners can be trusted to cooperate (Dawes, 1980). Researchers employ simplified interactions, such as the trust game (Fig. 1), to capture the essential structure of this dilemma (Berg et al., 1995). In the binary version of the game (Snijders & Keren, 1999), the *trustor* can decide to keep the status quo, which ends the interaction, or to trust the other player. In the event of trust, the *trustee* can choose between betrayal and reciprocation. Reciprocity leads to equal payoffs for both, and these outcomes are better than the status quo. The trustee gains even more by choosing betrayal, but this leads to the worst possible outcome for the trustor. Trust is risky; once it is chosen, the trustee has full control over the outcomes, and is faced with the temptation to choose betrayal.

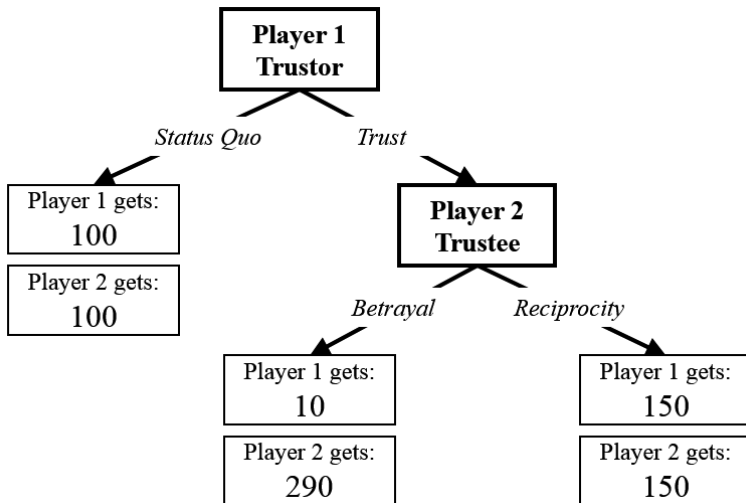


Figure 6.1. The sequential trust game with exemplary payoffs.

In the trust game, people rely on facial cues when making decisions (Ewing, Caulfield, Read, & Rhodes, 2014; Rezlescu et al., 2012; van't Wout & Sanfey, 2008). In general, trustworthy-looking partners are more likely to be trusted. Similar effects have been observed for consequential trust decisions outside the lab: Trustworthy-looking individuals ask for higher rent on Airbnb (Ert et al., 2016); have a higher chance of being granted loans on crowdsourcing websites (Duarte et al., 2012); and are more likely to be appointed as CEOs after firm misconduct (Gomulya et al., 2017). Facial trustworthiness also influences legal decisions (Porter, ten Brinke, & Gustaw, 2010; see also Berry & Zebrowitz-McArthur, 1988). Wilson and Rule (2015, 2016) found that untrustworthy-looking criminals were more likely to receive the death penalty (as opposed to life in prison). In short, perceived facial trustworthiness has far-reaching consequences.

Despite widespread reliance on facial cues, evidence on whether people are able to accurately infer others' trustworthiness from their facial appearance is mixed (Bonnefon et al., 2015; Todorov, Funk, et al., 2015). Some studies point to modest accuracy (Bonnefon et al., 2013; De Neys et al., 2017; C. Lin et al., 2018; Slepian & Ames, 2015; Tognetti et al., 2013), whereas others provide evidence against it both on empirical (Efferson & Vogt, 2013; Rule et al., 2013) and theoretical grounds (McCullough & Reed, 2016; Vogt et al., 2013). For example, Todorov and Porter (2014) found substantial variation in judgments of trustworthiness across different photos of the same person, which speaks against the idea that there are stable cues signaling an underlying trait (see also Sutherland, Young, & Rhodes, 2016). More evidence is needed to determine when people are able to accurately discern trustworthiness from facial cues. However, the current literature suggests that this ability is limited at best, and people would often make better decisions by relying on other cues (Bonnefon et al., 2017; Olivola & Todorov, 2010b).

The observation that facial trustworthiness judgments affect so many consequential decisions (in spite of their poor accuracy) has led several researchers to propose that the influence of face judgments constitutes a bias that should be eliminated (e.g., Olivola, Funk, & Todorov, 2014; Porter et al., 2010; Wilson & Rule, 2015). However, the origin of such a face bias (i.e., *why* there are such persistent effects of facial trustworthiness) remains poorly understood, and this shortcoming has hindered efforts aimed at curbing the bias. Here, we set out to address this gap.

### **Explanations for reliance on facial trustworthiness**

Scholars have long been interested in how people use different types of information, or cues, when making decisions (Brunswik, 1956; Slovic & Lichtenstein, 1971). Normative models propose that rational decision-makers should weigh all available cues according to how strongly they correlate with an outcome (Dawes, Faust, & Meehl, 1989). Yet, people often lack insight into the true validity of a cue (e.g., how diagnostic facial trustworthiness judgments really are) and descriptive decision models acknowledge that cues are weighed (or ranked) according to their *subjective*, rather than objective, validity (Brunswik, 1956; Gigerenzer & Goldstein, 1996; Hammond et al., 1964). A weaker definition of rationality holds that cue utilization should be determined by how valid people *think* the cue is. Reliance on facial trustworthiness may then be explained by the subjective belief that it is a particularly valid cue. This argument has been raised in a review by Olivola and colleagues (2014), who point to the historically persistent belief in the correspondence between facial cues and personality (Aristotle, trans. 1936; Lavater, 1775). For example, Hassin and Trope (2000) report that most people believe that faces contain at least some valid cues to individual personality traits (see also Suzuki, Tsukamoto, & Takahashi,

2017). We refer to this explanation of the face bias as the *subjective validity account*.

An alternative explanation can be derived from research into heuristic decision-making. People are often unable or unwilling to consider all available information and therefore rely on heuristics—strategies that consider only a subset of all available information—to reduce cognitive effort (Gigerenzer et al., 2011; Payne, Bettman, & Johnson, 1988; Simon, 1955). In fact, people favor cues that are intuitively available (Dimov & Link, 2017; Shah, 2007; Simmons & Nelson, 2006). Faces attract attention (Ro et al., 2001; Theeuwes & Van der Stigchel, 2006) and trustworthiness impressions are formed spontaneously (Klapper et al., 2016), quickly (Willis & Todorov, 2006), and effortlessly (Bonnefon et al., 2013). From this perspective, reliance on facial trustworthiness may be explained by the intuitive accessibility of the cue (Olivola & Todorov, 2010a; Willis & Todorov, 2006). We refer to this explanation of the face bias as the *intuitive accessibility account*.

## Overview of the studies

We present six studies to test the two accounts of the face bias: Do people persistently rely on facial trustworthiness because they believe it is a particularly valid cue, or because face judgments are intuitively accessible? We examine this question in the context of the trust game (Figure 6.1, Berg et al., 1995; Snijders & Keren, 1999), by comparing the effects of facial trustworthiness and economic payoff information, a cue we predict to be more subjectively valid, but also more difficult to evaluate. The combination of these two cues, which differ in subjective validity and accessibility, offers a critical test of the factors underlying cue preferences.

In Study 6.1, we test the perceived validities of facial trustworthiness and economic payoff information, with the prediction that people believe economic payoffs are more valid cue than facial

trustworthiness. In Study 6.2, we examine the time it takes to make decisions based on facial trustworthiness and economic payoffs; here, we expect it takes less time to reach a decision based on facial trustworthiness than on economic payoffs. Studies 6.3-6.5 examine how much people rely on the two cues when both are presented simultaneously. In these studies, we determine whether people favor the subjectively more valid cue (economic payoffs), or the cue that is processed more efficiently (facial trustworthiness). Finally, Study 6.6 tests how cognitive reflection affects reliance on the two cues. If people favor facial trustworthiness because it is easier to process than payoff information, then making intuitive (vs. reflective) decisions should reduce reliance on economic payoffs, but not reduce reliance on facial trustworthiness.

All data and analysis scripts are available at the Open Science Framework (<https://osf.io/h6dsj/>).

### **Study 6.1**

Study 6.1 examined explicit preferences for facial trustworthiness and economic payoff information. Evans and Krueger (2011, 2014) found that trustors are aware of the importance of the trustee's economic payoffs—people trust less when the trustee faces a greater temptation (i.e., a greater economic incentive to choose betrayal). In turn, trustees actually reciprocate less when temptation is large (Evans & Krueger, 2014). Thus, like facial trustworthiness, temptation is used as a cue when making trust decisions. However, prior work has not examined which of the two cues is seen as more valid. Given that temptation is actually predictive of trustees' behavior while facial trustworthiness has poor predictive validity, we set out to test whether people's explicit preferences for the two cues correspond to their predictive validities. To address this question, we let participants play a trust game where participants could choose which cue they would want



to have available: temptation (the trustee's economic payoff to betray trust) or facial trustworthiness (a facial photograph of their interaction partner).

## Method

**Participants.** Dutch undergraduate psychology students ( $N = 126$ ; 81.0% female) from Tilburg University participated in the study in return for partial course credit ( $M_{\text{age}} = 19.83$ ,  $SD_{\text{age}} = 2.96$ ). The sample size was based on the number of students that participated in the study within one week.

**Materials and procedure.** The experiment was administered in the lab. First, participants saw an exemplary decision tree (see Figure 6.1) and learned about the rules of the trust game. They were told that they would have to make a single decision between IN and OUT (trust and status quo) and that they would be randomly paired with another participant who would act as the trustee. If they chose OUT, the interaction would end. If they chose IN, then the trustee would make a choice between RIGHT and LEFT (reciprocation and betrayal). Choices were hypothetical and participants were asked to imagine that the points they were playing for would be converted into actual money at the end of the game. All participants played the role of the trustor.

Next, participants chose what kind of information (i.e., economic payoff or facial appearance) they wanted to have available. They could choose to see *how much your interaction partner would gain in case he/she chooses LEFT* or *a photo of your interaction partner*. The order in which they read the description of the two cues was counterbalanced. Participants then saw a different decision tree (with different payoffs) that was relevant for their choice. Depending on their choice of cue, either the betrayal payoff or the photo of their partner was revealed. The photo was taken from the Radboud Faces Database (RaFD; Langner et al., 2010) and showed a female person with a happy facial expression.

The trustee's temptation corresponded to a 100% increase in payoffs when choosing betrayal over reciprocation. Participants made their one-shot trust decision and indicated their confidence in having made the right decisions by dragging a slider along a scale ranging from 0 (*not at all confident*) to 100 (*extremely confident*).

### Results and discussion

To measure preferences for the two types of cues, we compared the percentages of participants who chose to see the photo of their partner versus the payoffs of their partner. A clear majority of participants (78.6%) preferred to see the trustee's temptation,  $p < .001$  (exact binomial test). Participants explicitly prioritized economic payoff information over facial trustworthiness information.

After cue selection, a total of 23.02% of participants trusted. Confidence ratings were similar for trust ( $M = 58.45$ ,  $SD = 18.85$ ) and distrust decisions ( $M = 62.92$ ,  $SD = 23.11$ ),  $t(55.54) = 1.06$ ,  $p = .29$ ,  $d = 0.22$ .<sup>24</sup> There was also no significant difference in confidence ratings between participants who chose to see their partner's payoffs ( $M = 61.86$ ,  $SD = 22.80$ ) and those who chose to see their partner's photo ( $M = 62.00$ ,  $SD = 20.30$ ),  $t(45.53) = 0.03$ ,  $p = .98$ ,  $d = 0.01$ . Confidence was not affected by the type of information available at the moment of decision-making.

### Study 6.2

In our first study, people perceived economic payoff information to be more valid than facial trustworthiness information. Our second study had two goals: First, we wanted to compare the cognitive effort required to make decisions based on payoffs versus facial trustworthiness.

---

<sup>24</sup> We report the results of Welch's  $t$ -tests which, compared to a Student's  $t$ -tests, provides equal power but superior error control in case of unequal variances between groups (Delacre, Lakens, & Leys, 2017).

Previous research has shown that facial trustworthiness is processed spontaneously, quickly, and effortlessly (Bonnefon et al., 2013; Klapper et al., 2016; Willis & Todorov, 2006). However, it is still unclear if making decisions based on facial trustworthiness is less effortful than making decisions based on other cues. Reduced decision effort should be reflected in faster decision times (cf. Bettman, Johnson, & Payne, 1990). We therefore hypothesized that people would make decisions more quickly when they relied on facial trustworthiness rather than economic payoffs. Differences in response times may also be caused by differences in decision conflict rather than decision effort (Evans & Rand, 2019; Krajbich, Bartling, Hare, & Fehr, 2015). Thus, we measured participants' confidence, as increased decision conflict leads to decreased confidence (De Neys, Cromheeke, & Osman, 2011; Zakay, 1985).

Second, we examined how facial and economic cues influenced the extent to which people rely on expectations of reciprocity. Expectations play a central role in psychological and economic models of trust, providing a conceptual link between cues and trust decisions (Bacharach & Gambetta, 2001; Thielmann & Hilbig, 2015). People are more likely to trust if they have high expectations of reciprocity (Costa-Gomes, Huck, & Weizsäcker, 2014; Thielmann & Hilbig, 2014). The subjective validity account suggests that people should rely more on their expectations when they are based on economic payoffs (vs. facial trustworthiness) since economic payoffs are subjectively more valid. The intuitive accessibility account makes the opposite prediction, as using appearance to form expectations requires less cognitive effort. Thus, comparing the extent to which people rely on their cue-based expectations served as a critical test of the two accounts.

## Method

**Participants.** A total of 134 students from Tilburg university (75.4% female;  $M_{\text{age}} = 21.30$ ,  $SD_{\text{age}} = 1.45$ ) participated in exchange for

partial course credit. The sample size was based on the number of students that participated in the study within two weeks.

**Materials and procedure.** The experiment was administered online. Participants were randomly assigned to the temptation condition or the face condition. In both conditions, participants first learned about and then played a series of 24 hypothetical trust games in the role of the trustor.

In the face condition, participants saw a photo of their interaction partner next to the decision tree on each trial. The photos were again taken from the RaFD (Langner et al., 2010). We selected 24 frontal photos of Caucasian Dutch adults with a forward gaze, of which half were male and half were female. Similar to previous investigations, half of the selected faces displayed a neutral expression and half a happy facial expression (i.e., they were smiling) in order to introduce variance in the perceived trustworthiness of the faces (cf. Evans & van de Calseyde, 2017). Previous research has shown that smiling individuals are perceived to be more trustworthy (Krumhuber et al., 2007; Said et al., 2009). To ensure that any observed effect of facial trustworthiness is not due to the attractiveness of the face (R. K. Wilson & Eckel, 2006), we selected faces of 12 male and 12 female individuals judged to be equally attractive (Langner et al., 2010).

In the temptation condition, we varied the trustee's temptation, i.e., the economic incentive to choose betrayal. Following previous work by Evans and Krueger (2014), we defined temptation as the difference between the trustee's gain in case of betrayal ( $T$ ) and reciprocation ( $R_2$ ) divided by the value of betrayal:  $(T - R_2)/T$  (see Figure 6.1). On half of the trials, temptation was low (0.33) and on the other half, it was high (0.60). These values correspond to a 50% (low temptation) and 150% (high temptation) increase in payoffs for the trustee in case betrayal is chosen over reciprocation. Note that payoffs in the face condition always corresponded to a temptation parameter of 0.5 (100% increase in

payoffs) which ensured that the average temptation across the 24 trials was equal in both conditions.

We assessed participants' response time for each trust game decision. Five extremely slow responses (0.16% of all decisions) were excluded because they were between 3 and 56 standard deviations slower than the mean. Response times were  $\log_{10}$ -transformed to account for their right-skewed distribution.

After making decisions in the 24 trust games, participants were shown each trust game again and they were asked to state their expectations of reciprocity. Specifically, we asked: *How likely is it that Player 2 will choose RIGHT [reciprocation]?* They could drag a slider along a scale ranging from 0 (*Player 2 will definitely choose LEFT [betrayal]*) to 100 (*Player 2 will definitely choose RIGHT [reciprocation]*). Participants also rated how confident they are that their expectations are accurate by dragging a slider along a scale ranging from 0 (*I am not confident at all*) to 100 (*I am extremely confident*). Following this, participants in the face condition saw each face one more time and were asked to rate how trustworthy the person in the photo is. Again, they could drag a slider along a scale ranging from 0 (*not trustworthy at all*) to 100 (*extremely trustworthy*). Similar to previous work, we used the average trustworthiness rating of each face across all participants as our measure of facial trustworthiness (van't Wout & Sanfey, 2008).<sup>25</sup>

## Results

**Descriptive statistics.** Average trustworthiness ratings of the faces ranged from 33.80 to 69.45 ( $M = 51.10$  out of 100,  $SD = 11.78$ ) and participants showed significant consensus in their ratings of the faces,  $ICC = .314$ ,  $p < .001$ , 95% CI [.204, .499]. Overall, participants trusted on

---

<sup>25</sup> In Study 6.2 and Study 6.3, we also measured individual differences in risk-taking, what participants wanted to do, and what they thought they should do for exploratory purposes.

37.47% of all trials and it took them on average 5.53 seconds ( $SD = 6.69$ ) to reach a decision. Sixteen participants (11.94%) never trusted whereas none of our participants always trusted. Average expectations of reciprocity were below the scale midpoint (50), indicating that participants believed that, across all trials, trustees were more likely to betray than to reciprocate trust ( $M = 43.23$ ,  $SD = 16.49$ ),  $t(133) = 4.75$ ,  $p < .001$ .

**Response times.** Making the same decision on all trials may lead to faster response times and ultimately obscure how response times are related to cue reliance. Results indeed showed that the sixteen participants who never trusted made significantly faster decisions ( $M = 0.394$ ,  $SD = 0.326$ ) than participants whose trust decisions varied across trials ( $M = 0.602$ ,  $SD = 0.200$ ),  $t(16.57) = 2.49$ ,  $p = .024$ ,  $d = 0.66$ . We therefore excluded these participants from our response time analyses.

We compared the time participants took to make decisions based on temptation versus facial trustworthiness. We hypothesized that decisions based on facial trustworthiness would be faster than decisions based on temptation. A  $t$ -test comparing response times between the two conditions showed that participants in the face condition reached their decisions substantially faster ( $M = 0.437$ ,  $SD = 0.198$ ) than participants in the temptation condition ( $M = 0.709$ ,  $SD = 0.167$ ),  $t(125.4) = 8.58$ ,  $p < .001$ ,  $d = 1.58$  (see Figure 6.2A).

We also correlated each participant's average response time with the extent to which they relied on the available cue *within* each condition.<sup>26</sup> To test the hypothesis that reliance on temptation is more effortful than reliance on facial cues, we regressed participants' average response times on their level of cue reliance, our condition variable (coded -0.5 for the temptation condition and 0.5 for the face condition),

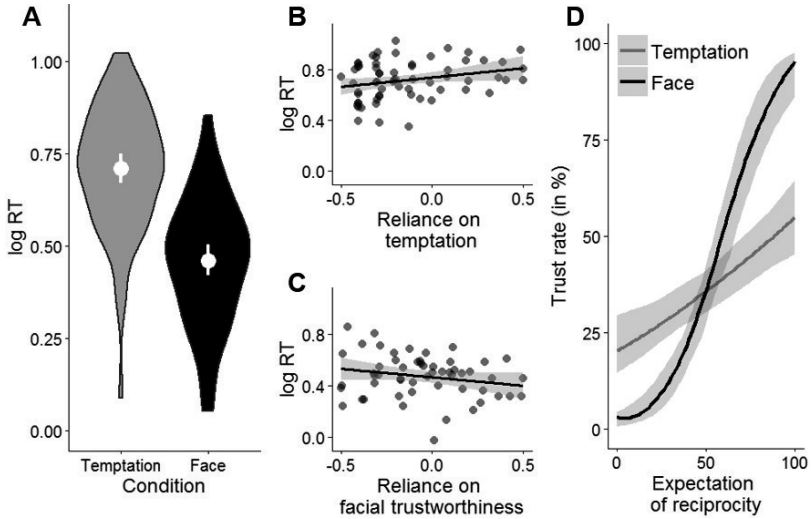
---

<sup>26</sup> We extracted the absolute effects of temptation and facial trustworthiness (depending on the condition) on trust decisions from our multilevel regression models as an indication of the extent to which each participant relied on the cue. Like all other predictors, we standardized this cue reliance variable.

and an interaction term of the two variables. Response times were faster in the face condition,  $b = -0.272$ ,  $SE = 0.031$ ,  $p < .001$ , and cue reliance (across both conditions) had no significant effect on response times,  $b = 0.008$ ,  $SE = 0.053$ ,  $p = .89$ . More importantly, we found a significant interaction effect between cue reliance and condition,  $b = -0.278$ ,  $SE = 0.106$ ,  $p = .010$ .

To understand this interaction, we examined the correlation between cue reliance and average response time within each experimental condition. In the temptation condition, we found a positive correlation between the extent to which participants' relied on temptation and the average time needed to make a decision,  $r(61) = .265$ ,  $p = .036$  (see Figure 6.2B). In the face condition, we found no significant correlation between the extent to which participants relied on facial trustworthiness and their response times,  $r(53) = -.214$ ,  $p = .12$  (see Figure 6.2C). Increased reliance on temptation was associated with longer response times, providing further evidence that reliance on temptation is more effortful than reliance on facial trustworthiness.

Response time differences between the two conditions may also be driven by differences in decision conflict (Krajbich et al., 2015). To test this account, we compared participants' confidence ratings between the two conditions. If participants made slower decisions based on temptation because they experienced more decision conflict, they should also show decreased confidence (De Neys et al., 2011; Zakay, 1985). On average, participants in the face condition ( $M = 60.51$ ,  $SD = 14.38$ ) were not more confident than participants in the temptation condition ( $M = 64.45$ ,  $SD = 14.18$ ),  $t(131.28) = 1.60$ ,  $p = .11$ ,  $d = 0.28$ . We also found no evidence that confidence was related to how much people relied on temptation,  $r(67) = -.154$ ,  $p = .21$ , or on facial trustworthiness,  $r(63) = -.124$ ,  $p = .32$ . Thus, results showed no evidence that response time differences between the two conditions were due to differences in decision conflict.



*Figure 6.2.* The effects of facial cues and payoff information on response times and reliance on expectations (Study 6.2): (A) Violin plots showing the difference in  $\log_{10}$ -transformed response times between the temptation condition and the face condition. Dots denote the mean and bars denote the 95% confidence interval; (B) the correlation between participants' reliance on temptation and their average response times; (C) the correlation between participants' reliance on facial trustworthiness and their average response times; (D) the relationship between expectations of reciprocity and the probability of trust when expectations were based on temptation vs. facial trustworthiness. Values denote the predicted probability of trust derived from multilevel regression models.

**Reliance on expectations of reciprocity.** To conclude, we examined how expectations of reciprocity influenced participants' trust decisions. We estimated multilevel regression models including random intercepts and random slopes. Regressing trust decisions on expectations, condition, and an interaction term of the two variables yielded a positive effect of expectations,  $b = 0.038$ ,  $SE = 0.004$ ,  $p < .001$ ,  $OR = 1.04$ , and a negative effect of condition,  $b = -2.095$ ,  $SE = 0.450$ ,  $p < .001$ ,  $OR = 0.12$ . More importantly, we found a significant interaction effect between expectations and condition,  $b = 0.042$ ,  $SE = 0.009$ ,  $p < .001$ ,



$OR = 1.04$ .<sup>27</sup> To probe this interaction effect, we examined the effects of expectations within each condition. In line with the intuitive accessibility account, the effect of expectations on trust was larger in the face condition,  $b = 0.067$ ,  $SE = 0.010$ ,  $p < .001$ ,  $OR = 1.07$ , than in the temptation condition,  $b = 0.015$ ,  $SE = 0.004$ ,  $p < .001$ ,  $OR = 1.02$  (see Figure 6.2D). Thus, participants relied more on their expectations of reciprocity when these could be formed based on the cue that is more intuitively accessible, rather than based on the cue that is seen as more valid.

## Discussion

The results of Study 6.2 shed more light on how facial trustworthiness and economic payoffs differently influence trust decisions. Our results extend previous findings on the processing of facial trustworthiness (e.g., Bonnefon et al., 2013; Klapper et al., 2016; Willis & Todorov, 2006) by showing that making decisions based on facial trustworthiness requires less cognitive effort than making trust decisions based on payoffs. Trust decisions based on payoffs took longer than trust decisions based on facial trustworthiness, and reliance on economic payoffs, but not facial trustworthiness, was positively correlated with decision time.

We also found that the expectations of reciprocity were positively correlated with trusting behavior. Expectations influenced trust decisions when they were based on either payoffs or facial trustworthiness. Crucially, consistent with the intuitive accessibility account, participants relied more on their expectations when they were

---

<sup>27</sup> We also examined the possibility of a non-linear relationship between expectations and trust decisions by estimating regression models that included linear, quadratic, and cubic terms for the effect of expectations. A comparison of model fits showed that the linear model provided the best fit. Adding a quadratic term to the model did not significantly increase model fit ( $p = .17$ ) and neither did adding a cubic term ( $p = .89$ ).

based on facial trustworthiness (even though Study 6.1 showed that facial trustworthiness is seen to be a less valid cue). We suggest that this is due to the fact that forming expectations based on easily accessible face judgments is less effortful than considering economic payoffs.

### **Studies 6.3-6.5**

Our next studies were designed to test predictions of the subjective validity and intuitive accessibility accounts more directly. We examined how the presence of facial trustworthiness information affects reliance on economic payoff information (Study 6.3). The intuitive accessibility account predicts that people will rely less on economic payoffs when they can also rely on facial trustworthiness, since it takes less effort to rely on the latter. On the other hand, the subjective validity account predicts that how much people rely on payoffs will not depend on whether or not facial trustworthiness is available, since economic information is seen as more valid.

We also examined how the presence of economic payoff information influenced reliance on facial trustworthiness (Study 6.4). The intuitive accessibility account predicts that how much people rely on facial trustworthiness will not depend on whether or not economic payoff information is also available. On the other hand, the subjective validity account predicts that people will rely less on facial trustworthiness information when they can also rely on economic payoff information, since the latter is seen as a more valid cue. This setup also allowed us to address an alternative explanation for the discounting of economic information. It is plausible that any cue is discounted if another cue (that is seen as at least somewhat valid) is available as well. Based on this alternative explanation, one would also expect people to discount facial trustworthiness in the presence of economic payoff information.

After conducting these two initial studies, we ran a third pre-registered study that examined both reliance on temptation and facial

trustworthiness in an integrated design (Study 6.5). We report aggregated results of the three studies since they examined the same hypotheses with almost identical designs using participants from the same source.

## Method

**Participants.** We recruited a total of 2,007 (Study 6.3:  $n = 201$ , Study 6.4:  $n = 200$ , Study 6.5:  $n = 1,606$ ) U. S. American workers from Amazon Mechanical Turk (MTurk; Paolacci & Chandler, 2014) who participated in exchange for \$1. Participants who failed an attention check at the end of the study or who indicated having only a poor or basic English proficiency (Study 6.3:  $n = 22$ , Study 6.4:  $n = 18$ , Study 6.5:  $n = 292$ ) were excluded from analysis leaving a final sample of 1,675 participants (52.37% female,  $M_{\text{age}} = 34.98$ ,  $SD_{\text{age}} = 10.75$ ).

**Materials and procedure.** The experiments were administered online. Similar to Study 6.2, all participants learned the rules of the trust game and made a series of 24 (Study 6.3 and 6.4) or 32 (Study 6.5) hypothetical trust game decisions. In Study 6.3, participants were randomly assigned to the temptation-only condition or the face-and-temptation condition; in Study 6.4, participants were randomly assigned to the face-only condition or the face-and-temptation condition; and in Study 6.5, participants were randomly assigned to the temptation-only condition, the face-only condition, or the face-and-temptation condition.

In the temptation-only condition ( $n = 528$ ), trustee's temptation to betray was varied. On half of the trials, temptation was low (0.2) and on the other half, temptation was high (0.60). These values correspond to a 25% (low temptation) and 150% (high temptation) increase in payoffs for the trustee in case betrayal is chosen over reciprocation.

In the face-only condition ( $n = 525$ ), participants saw a photo of their interaction partner and the level of temptation was held constant. In Study 6.4, we selected twelve photos that were already used in Study

6.1 (six male and six female matched on attractiveness; three individuals with a happy and three with a neutral expression for each gender). In Study 6.5, we selected a total of sixteen photos that were already used in Study 6.1 (eight male and eight female matched on attractiveness; four individuals with a happy and four with a neutral expression for each gender). Participants interacted twice with each individual.

In the face-and-temptation condition ( $n = 622$ ), we varied both cues orthogonally. In Study 6.3, we selected four photos that were already used in Study 6.1 (two male and two female matched on attractiveness; one individual with a happy and one with a neutral expression for each gender). Each photo was presented six times—three times paired with low and three times paired with high temptation. In Study 6.4 and Study 6.5, each of the selected photos that were also displayed in the face-only condition was presented twice—once paired with low and once paired with high temptation.

Participants were then shown each face again and asked to rate how trustworthy they think the person in the photo is. The average trustworthiness rating of each face across all participants was used as our measure of facial trustworthiness. We removed 285 decisions (0.56% of all decisions) with extremely slow response times because they were between 3 and 78 standard deviations slower than the mean. Response times were  $\log_{10}$ -transformed to account for their right-skewed distribution.<sup>28</sup>

## Results

We rescaled our cue variables to range from -0.5 to 0.5. Thus, for our two cues, a one-unit increase denotes a change from low to high

---

<sup>28</sup> In Study 6.3 and Study 6.4, we again measured participants' expectations of reciprocity and their confidence in their expectations. For the sake of brevity, we report the results in the Supplemental Materials.

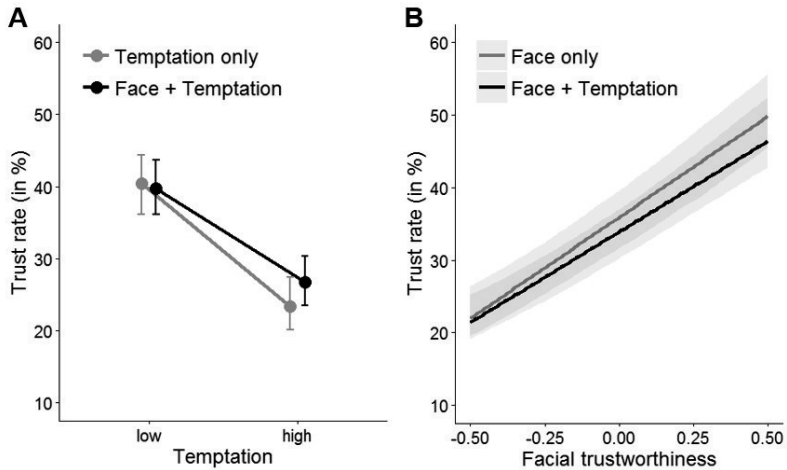
temptation and a change from the lowest average trustworthiness rating to the highest.

**Descriptive statistics.** Average trustworthiness ratings of the faces ranged, from 35.97 to 74.48 ( $M = 54.87$ ,  $SD = 10.92$ ). Participants showed significant consensus in their ratings of the faces in Study 6.3,  $ICC = .281$ ,  $p < .001$ , 95% CI [.104, .848], Study 6.4,  $ICC = .294$ ,  $p < .001$ , 95% CI [.170, .548], and Study 6.5,  $ICC = .294$ ,  $p < .001$ , 95% CI [.185, .500]. The average trust rate across all trials was 39.71% and the average response time was 3.82 seconds ( $SD = 3.79$ ). A total of 199 participants (11.88%) never trusted whereas 82 participants (4.90%) always trusted.

**Reliance on temptation.** First, we tested how much participants relied on temptation when it was the only available cue versus when both facial trustworthiness and temptation were available. To this end we estimated a multilevel regression model with random intercepts per participant and random slopes for all predictors, in which we regressed participants' trust decisions on temptation, condition (coded -0.5 for temptation-only and 0.5 for face-and-temptation), and an interaction term between the two variables. This yielded a negative effect of temptation,  $b = -0.688$ ,  $SE = 0.047$ ,  $p < .001$ ,  $OR = 0.50$ , and no effect of condition,  $b = 0.082$ ,  $SE = 0.119$ ,  $p = .49$ ,  $OR = 0.86$ . Crucially, we observed a significant interaction effect between temptation and condition,  $b = 0.198$ ,  $SE = 0.095$ ,  $p = .036$ ,  $OR = 1.22$  (see Figure 6.3A). In line with the intuitive accessibility account, but contrary to the subjective validity account, participants relied somewhat less on temptation when they could also rely on facial trustworthiness,  $b = -0.589$ ,  $SE = 0.056$ ,  $p < .001$ ,  $OR = 0.55$ , as opposed to when temptation was the only available cue,  $b = -0.787$ ,  $SE = 0.076$ ,  $p < .001$ ,  $OR = 0.46$ .

**Reliance on facial trustworthiness.** Next, we tested how much participants relied on facial trustworthiness when it was the only available cue versus when both facial trustworthiness and temptation

were available. To this end, we estimated a multilevel regression model with random intercepts per participant and random slopes for all predictors, in which we regressed participants' trust decisions on the facial trustworthiness of their interaction partner, condition (coded -0.5 for face-only and 0.5 for face-and-temptation), and an interaction term between the two variables. This yielded a positive effect of facial trustworthiness,  $b = 1.192$ ,  $SE = 0.071$ ,  $p < .001$ ,  $OR = 3.29$ , and no effect of condition,  $b = -0.082$ ,  $SE = 0.119$ ,  $p = .49$ ,  $OR = 0.92$ . Crucially, we did not observe a significant interaction effect between facial trustworthiness and condition,  $b = 0.107$ ,  $SE = 0.143$ ,  $p = .46$ ,  $OR = 0.90$  (see Figure 6.3B). In line with the intuitive accessibility account, but contrary to the subjective validity account, we found no evidence that participants relied less on facial trustworthiness when they could also rely on temptation,  $b = 1.139$ ,  $SE = 0.094$ ,  $p < .001$ ,  $OR = 3.12$ , as opposed to when facial trustworthiness was the only available cue,  $b = 1.245$ ,  $SE = 0.108$ ,  $p < .001$ ,  $OR = 3.47$ .



*Figure 6.3.* The discounting of economic payoff information in the presence of facial cues and vice versa (Studies 6.3-6.5): (A) The effect of temptation on trust rates when temptation was the only available cue vs. when both facial trustworthiness and temptation were available. (B) The effect of facial trustworthiness on trust rates when facial trustworthiness was the only available cue vs. when both facial trustworthiness and temptation were available. Values denote the predicted probability of trust derived from multilevel regression models.

**Response times.** We again examined participants' response times to test how long they took to make trust decisions when relying on temptation or facial trustworthiness. Contrary to the results of Study 6.2, a *t*-test comparing response times between the temptation-only and face-only condition showed no evidence that participants in the face-only condition made faster decisions ( $M = 0.473$ ,  $SD = 0.239$ ) than participants in the temptation-only condition ( $M = 0.444$ ,  $SD = 0.230$ ),  $t(880.47) = 1.85$ ,  $p = .064$ ,  $d = 0.12$ .

To test the relationship between cue reliance and response times more directly, we regressed participants' average response times on their reliance on temptation, condition (coded -0.5 for temptation-only and 0.5 for face-and-temptation), and an interaction term between the

two variables. This revealed a positive effect of condition,  $b = 0.046$ ,  $SE = 0.015$ ,  $p = .002$ , a positive effect of reliance on temptation,  $b = 0.040$ ,  $SE = 0.007$ ,  $p < .001$ , but no significant interaction effect between reliance on temptation and condition,  $b = 0.013$ ,  $SE = 0.014$ ,  $p = .34$ . Participants took longer to reach a decision when both temptation and facial trustworthiness were varied compared to when only temptation was varied. More importantly though, as in Study 6.2, we observed that the more participants relied on temptation, the longer they took to decide, and this effect did not significantly vary between the two conditions.

Similarly, we regressed participants' average response times on their reliance on facial trustworthiness, condition (coded -0.5 for face-only and 0.5 for face-and-temptation), and an interaction term between the two variables. This revealed no effect of condition,  $b = 0.010$ ,  $SE = 0.015$ ,  $p = .53$ , no effect of reliance on facial trustworthiness,  $b = -0.008$ ,  $SE = 0.007$ ,  $p = .25$ , and no interaction effect between reliance on facial trustworthiness and condition,  $b = 0.001$ ,  $SE = 0.014$ ,  $p = .94$ . Thus, as in Study 6.2, we found no evidence that an increased reliance on facial trustworthiness was related to longer response times.

## Discussion

In sum, participants still relied on facial trustworthiness in the presence of another (more subjectively valid) cue (i.e., economic payoffs). More importantly, the presence of facial cues led participants to rely somewhat *less* on the subjectively more valid cue. This pattern of results is in line with the notion that people rely on facial trustworthiness because it is relatively quick and effortless.

## Study 6.6

The results so far suggest that people favor relying on facial trustworthiness because it is intuitively accessible. Response time data from Study 6.2 and Studies 6.3-6.5 provided support for this argument,



as increased reliance on economic payoffs was related to longer decision times (while there was no relationship between reliance on facial cues and decision times). The goal of Study 6.6 was to provide experimental evidence for this claim by testing how reliance on economic payoffs and facial trustworthiness varies when participants make intuitive (vs. reflective) decisions. If reliance on economic payoffs requires cognitive reflection, we would expect an attenuated effect of payoff information when people make intuitive (vs. reflective) decisions.

Regarding reliance on facial trustworthiness, two predictions are plausible: People may override their trait impressions and rely more on economic information when making reflective as opposed to intuitive decisions, as they deem economic information to be the more valid cue. Alternatively, people may not be fully aware of how their decisions are influenced by their intuitive face judgments, or they may be unable to suppress or correct the influence of face judgments (T. D. Wilson & Brekke, 1994). From this perspective, additional reflection would not necessarily undermine the effect of face judgments on decisions.

## Method

**Participants.** We recruited a sample of 962 U. S. American MTurk workers who participated in exchange for \$2 each. Data from participants who failed an attention check at the end of the survey or who indicated having only a poor or basic English proficiency were excluded from analysis, leaving a final sample of 797 participants (47.55% female,  $M_{\text{age}} = 36.08$ ,  $SD_{\text{age}} = 11.31$ ).

**Materials and procedure.** The experiment was administered online. All participants learned the rules of the trust game and made a series of 24 hypothetical decisions, during which their interaction partners' temptation to betray and their facial trustworthiness were varied. We used the same twelve photos that were already used in Study 6.4 (six male and six female matched on attractiveness; three individuals

with a happy and three with a neutral expression for each gender). Participants interacted twice with each individual—once when temptation was low (0.2; 25% gain from betrayal) and once when temptation was high (0.60; 150% gain from betrayal). Participants were randomly assigned to one of two conditions.

In the intuition condition ( $n = 399$ ), participants were prompted to follow their first instinct and make intuitive decisions (adapted from Newman, Gibb, & Thompson, 2017). We asked participants to reach each decision within five seconds. A timer counting backwards from five to zero was displayed on each decision page. Participants could still indicate a decision after the timer reached zero and the page only forwarded to the next trial once a decision was made. In the reflection condition ( $n = 398$ ), participants were prompted to think carefully and make reflective decisions. Participants were informed that on each trial, they could only indicate a decision after ten seconds had passed. They were asked to take at least ten seconds to weigh all options and reflect on their decision. After they had made their decisions, participants were shown each face again and we asked them to rate how trustworthy they think the person in the photo is. We used the average trustworthiness rating of each face across all participants as our measure of facial trustworthiness.

## Results

**Descriptive statistics.** Average trustworthiness ratings of the faces ranged from 36.35 to 74.24 ( $M = 55.25$ ,  $SD = 13.19$ ) and participants showed significant consensus in their ratings of the faces,  $ICC = .372$ ,  $p < .001$ , 95% CI [.229, .631]. The average trust rate across all trials was 44.99% ( $SD = 25.44\%$ ) in the intuition condition and 37.32% ( $SD = 25.45\%$ ) in the reflection condition. The difference in trust rates between the two conditions was significant,  $t(795) = 4.25$ ,  $p < .001$ ,  $d = 0.30$ . A total of 91 participants (11.42%) never trusted whereas 26

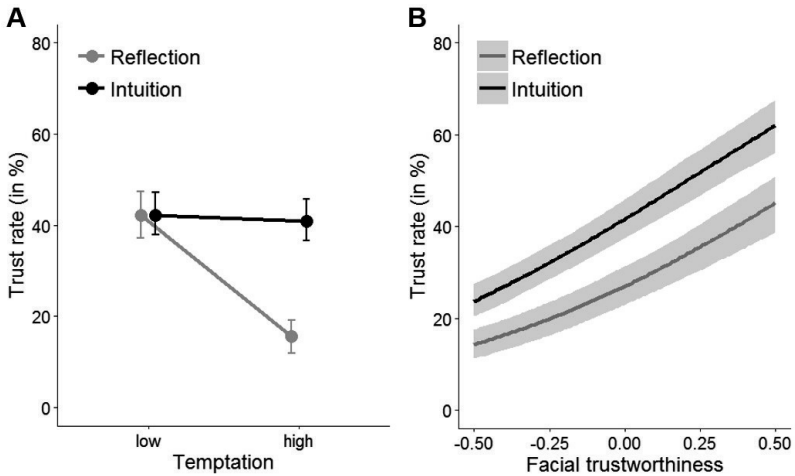
participants (3.26%) always trusted. On average, participants in the intuition condition took 2.40 seconds to make a decision ( $SD = 4.04$ ), whereas participants in the reflection condition took 13.66 seconds ( $SD = 0.98$ ),  $t(481.9) = 70.97$ ,  $p < .001$ ,  $d = 5.03$ .

**Reliance on temptation.** We first tested the prediction that participants would rely less on payoff information when making intuitive (vs. reflective) decisions. We estimated a multilevel regression model with random intercepts per participant and random slopes for all predictors, in which we regressed participants' trust decisions on temptation, facial trustworthiness, condition (coded -0.5 for reflection and 0.5 for intuition), and an interaction term between temptation and condition. This yielded a negative effect of temptation,  $b = -0.670$ ,  $SE = 0.059$ ,  $p < .001$ ,  $OR = 0.45$ , a positive effect of facial trustworthiness,  $b = 1.607$ ,  $SE = 0.104$ ,  $p < .001$ ,  $OR = 4.99$ , and a positive effect of condition,  $b = 0.547$ ,  $SE = 0.136$ ,  $p < .001$ ,  $OR = 1.73$ . Crucially, we observed a significant interaction effect between temptation and condition,  $b = 1.185$ ,  $SE = 0.115$ ,  $p < .001$ ,  $OR = 3.27$  (see Figure 6.4A). Participants relied less on temptation when making intuitive decisions,  $b = -0.055$ ,  $SE = 0.059$ ,  $p = .35$ ,  $OR = 0.95$ , than when making reflective decisions,  $b = -1.376$ ,  $SE = 0.112$ ,  $p < .001$ ,  $OR = 0.25$ .

**Reliance on facial trustworthiness.** Next, we tested how much participants relied on facial trustworthiness when making intuitive (vs. reflective) decisions. We again estimated a multilevel regression model, this time including an interaction term between facial trustworthiness and condition. This did not yield a significant interaction effect,  $b = 0.237$ ,  $SE = 0.218$ ,  $p = .27$ ,  $OR = 1.27$  (see Figure 6.4B). Participants relied on facial trustworthiness when making intuitive decisions,  $b = 1.660$ ,  $SE = 0.153$ ,  $p < .001$ ,  $OR = 5.26$ , and when making reflective decisions,  $b = 1.596$ ,  $SE = 0.142$ ,  $p < .001$ ,  $OR = 4.93$ .<sup>29</sup>

---

<sup>29</sup> Excluding 15 participants (1.88%), who indicated not having made intuitive or reflective decisions on all trials did not change the pattern of results.



*Figure 6.4.* The effects of temptation and facial trustworthiness on trust rates when making intuitive versus reflective decisions (Study 6.6): (A) The effect of temptation on trust rates when participants made intuitive (vs. reflective) trust decisions. (B) The effect of facial trustworthiness on trust rates when participants made intuitive (vs. reflective) trust decisions. Values denote the predicted probability of trust derived from multilevel regression models.

## Discussion

Results of Study 6.6 revealed that the influence of temptation, but not facial trustworthiness, on participants' trust decisions was reduced when participants made intuitive as opposed to reflective decisions. This finding is consistent with prior research suggesting that reliance on facial trustworthiness does not require cognitive reflection (Bonnefon et al., 2013; De Neys et al., 2017; Mieth, Bell, & Buchner, 2016). The results of the current study also shed light on the question of how the intuitive accessibility of face judgments causes their persistent effects on decision-making. Given enough time or motivation to reflect on their decisions, people might realize that relying on different information will lead to better decisions. However, it might also be the case that the effect

of facial cues is more implicit: People could be unaware of how their decisions are influenced by intuitive face judgments, or they could be unable to suppress this influence (T. D. Wilson & Brekke, 1994). From this perspective, additional reflection would not necessarily undermine the effect of facial cues on decisions. The current results are in line with this latter view, as we did not find any evidence that the effect of facial trustworthiness is diminished when people made reflective as opposed to intuitive decisions.

### General discussion

We asked whether reliance on facial trustworthiness in social decision-making can be explained by beliefs in the diagnosticity of the cue (i.e., by subjective cue weighing) or by the fact that the cue is intuitively accessible which makes it relatively effortless to rely on it (i.e., by intuitive accessibility). Across six studies, we systematically tested how much people rely on facial trustworthiness and economic payoff information (i.e., the trustee's temptation to betray trust)—a cue that is perceived to be more valid but takes more effort to process. Across our studies, we find consistent support for the intuitive accessibility account.

In Study 6.1, we found that economic payoff information is seen as a more valid cue than facial information. Study 6.2 showed there was a positive correlation between reliance on economic payoffs and decision time, while no such relationship was found for reliance on facial trustworthiness. These findings replicated in Studies 6.3-6.5, with a substantially larger sample. To the extent that longer response times are an indicator for more effortful decision strategies (Bettman et al., 1990), these results suggest that it takes less cognitive effort to rely on facial trustworthiness rather than temptation. Thus, Study 6.1 identified economic payoff information as a cue that is seen as more valid, whereas Study 6.2 and Studies 6.3-6.5 showed that economic payoff information is also more effortful to process than facial trustworthiness.

In Studies 6.3-6.5, we tested to what extent people rely on economic payoff information and facial trustworthiness when both cues are available simultaneously. Results showed that people rely less on a subjectively more valid cue (i.e., temptation) when facial trustworthiness is available as well. However, people do not rely less on facial trustworthiness when temptation is simultaneously manipulated. This pattern of results is in line with the argument that persistent reliance on facial trustworthiness is driven by the intuitive accessibility of the cue. If cognitive reflection is required for reliance on economic payoff information, but not for reliance on facial cues, then we would expect that restricting reflection during decision-making only decreases the influence of payoff information, but not the influence of facial cues. This prediction was explicitly tested in Study 6.6. This final study confirmed that payoff information had a diminished effect on trust when people made intuitive (vs. reflective) decisions. Taken together, our data suggest that the persistent reliance on facial trustworthiness can be better explained by the intuitively accessibility of the cue, which makes reliance on it relatively effortless, rather than by the belief that it is a particularly valid cue.

### **Decision effort and trust**

The present results converge with prior research on the processing of trait information from faces. A host of studies has demonstrated that trustworthiness impressions from faces are formed in a particularly efficient manner (Bonnefon et al., 2013; Klapper et al., 2016; Willis & Todorov, 2006). We extend these findings by showing that reliance on economic payoff information requires more effort than reliance on facial cues. Similar results were observed when we manipulated how much people could reflect on their decisions. Intuitive decisions were less reliant on economic payoff information, but the manipulation of reflection time had no effect on reliance on facial cues. Our interpretation

that reliance on facial trustworthiness is relatively effortless is also in line with studies showing that subjecting participants to a cognitive load manipulation does not impair their use of facial cues when making trust decisions (Bonnefon et al., 2013; De Neys et al., 2017; Mieth et al., 2016).

In Study 6.2 and Study 6.3-6.5, we measured response times as a proxy for the cognitive effort needed to make a decision (Bettman et al., 1990). Previous investigations in the domain of social decision-making have highlighted that other factors such as decision conflict can also drive response times (Evans, Dillon, & Rand, 2015; Krajbich et al., 2015). Perhaps, decision-makers felt more conflicted when relying on temptation, which increased their decision times. According to this view, we would also expect participants to be less confident when relying on temptation as decision conflict is associated with decreased confidence (De Neys et al., 2011; Zakay, 1985). Yet, across our studies, we found no evidence that participants reported lower levels of confidence in the presence of temptation as opposed to facial trustworthiness, suggesting that decision conflict was not driving our results.

Our results also fit within broader frameworks of how people decide whom to trust. With regard to the question which cues people rely on in trust situations, Evans and Krueger (2016) suggest in their model of bounded prospection that people approach trust decisions egocentrically: People focus on their potential costs and benefits while neglecting the probability that these outcomes will occur as assessing an interaction partner's trustworthiness requires perspective-taking and thus cognitive effort (S. Lin, Keysar, & Epley, 2010). We go beyond contrasting cues pertaining to the self with cues pertaining to the situation (Thielmann & Hilbig, 2015) or the interaction partner (Evans & Krueger, 2011) and provide evidence for a more general claim that the extent to which people rely on information in trust decisions with multiple cues is determined by the ease with which the information is processed.

### **Limitations and future research**

We provided evidence that the persistent reliance on facial trustworthiness can at least partly be explained by the intuitive accessibility of face judgments. We demonstrated this by showing that people favor relying on facial trustworthiness over temptation even though the latter is seen as a more valid cue. More specifically, we found that people did not rely less on facial trustworthiness when they could also rely on temptation. This finding stands in contrast to previous studies showing that people discount facial cues when other information is available (Graham et al., 2017; Rezlescu et al., 2012; Yu, Saleem, & Gonzalez, 2014). For example, Rezlescu and colleagues (2012) showed that, the influence of facial trustworthiness on trust decisions was reduced, but not eliminated, when information on a partner's past behavior was available. Maybe the subjective validity of temptation is still comparatively low and a stronger cue would have reduced or even eliminated reliance on facial cues in our studies. Arguably, people may prioritize past behavior over facial trustworthiness because past behavior is both easy to process and subjectively valid. To date, there is relatively little work that examines how people prioritize different types of cues in dilemmas of trust; future studies need to test how reliance on facial trustworthiness changes in the presence of different cues that vary in both subjective validity and ease of accessibility.

In addition, more research is needed to understand how the efficient processing of faces influences cue selection and cue weighing. Following the fast accessibility of face judgments, do people ignore other cues altogether? Our results speak against such a strong version of one-reason decision-making (Gigerenzer et al., 2008) as the effect of temptation was still apparent even when decision-makers could have relied solely on facial trustworthiness. Individual differences in rational versus heuristic processing (Epstein et al., 1996) or need for cognition (Cacioppo & Petty, 1982) might help explain why some individuals reach



a decision after face judgments are accessible whereas other consider additional cues. People who score higher on heuristic processing may rely more on facial trustworthiness even when other cues are available.

Alternatively, it could be the case that our participants did in fact consider all available cues but that the efficient processing of faces influenced how the cues were weighed. Shah (2007) has argued that efficiently processed cues are weighed more heavily. Since participants only read a description of the face cue and did not experience the efficient processing of it in Study 6.1, our measure of subjective cue validity might have underestimated how valid participants perceived the cue to be *while* making decisions. However, Dimov and Link (2017) showed that processing efficiency does not affect cue weights, but rather the order in which cues are considered. In the context of our study, it is thus conceivable that the discounting of temptation was due to a sequential process where fast facial trustworthiness judgments anchored participants' response and subsequent adjustment on the basis of temptation was insufficient (Tamir & Mitchell, 2012). This view holds that the primary processing of faces influences the weighing of subsequently processed cues. Future research could test whether a serial presentation of cues, where participants first learn about temptation and then see the face of their interaction partner, eliminates the discounting of temptation in favor of facial trustworthiness.<sup>30</sup>

---

<sup>30</sup> An anonymous reviewer raised the possibility that people might simply be reluctant to report that they see faces as valid cues. While more research is needed to measure how accurate people think their face judgments are, the available evidence suggests that people do think and report that faces contain information about a person's personality (Hassin & Trope, 2000; Suzuki et al., 2017). For example, in a survey by Hassin and Trope (2000), they found that 75% of respondents agreed that at least some traits can be read from a person's facial appearance.

### **Implications for debiasing interventions**

The primary goal of our current investigation was to shed light on the process giving rise to the face bias. In addition, our findings may also be used to inform the design of interventions aimed at curbing the bias. One previously proposed recommendation is to inform decision-makers about the poor accuracy of their impressions (Porter et al., 2010). Our results suggest that this approach might not be sufficient as the face bias does not stem from a conscious (albeit subjective) weighing of cues (Dawes et al., 1989; T. D. Wilson & Brekke, 1994). Another idea would be to increase the importance of the decision in order to motivate people to process the available information more deeply (Chaiken, 1980; Chaiken & Maheswaran, 1994) or to reduce processing constraints (e.g., time pressure) or decision complexity (Gigerenzer & Goldstein, 1996; Payne et al., 1988).

Both approaches might lead to an attenuation of the face bias but we are skeptical that it will completely eliminate reliance on facial cues. Previous studies have shown that the influence of facial trustworthiness persists for extremely important decisions, such as legal sentencing (Blair et al., 2004; Eberhardt et al., 2006; J. P. Wilson & Rule, 2015) or voting (Jaeger, Evans, & van Beest, 2019; Olivola & Todorov, 2010a), as well as for relatively simple, self-paced decisions in the lab (Mieth et al., 2016; Rezlescu et al., 2012). As a further case in point, Study 6.6 showed that, while making people reflect on their decision increased their reliance on temptation, we found no evidence that it reduced the influence of facial trustworthiness.

Our findings suggest that interventions that disrupt the primary and efficient processing of faces might hold more potential for success. Inverting photos or misaligning its parts might disrupt the efficient processing of faces while still enabling the identification of an individual (Bindemann, Burton, Hooge, Jenkins, & de Haan, 2005; Todorov, Loehr, & Oosterhof, 2010), but requesting photos to be displayed upside-down

in case files or on websites such as Airbnb is not realistic. However, in many situations, decision-makers could be asked to reach a decision in absence of facial cues first—ideally based on valid information (e.g., Jaeger, Todorov, Evans, & van Beest, 2019). In the next step, a photo is displayed while the decision can still be revised. With this setup, initial decisions should be less biased. Any incorporation of face judgments into the decision-making process would now require a conscious revision of the decision, which is less likely if more diagnostic cues are also available. Future studies should test interventions targeting the efficient processing of faces and compare their effectiveness to simply teaching decision-makers about the biasing potential of faces.

## Conclusion

Despite their poor predictive validity, face judgments influence decisions ranging from the relatively trivial choice of which apartment to rent (which might determine the fate of the next vacation; Ert et al., 2016), to the selection of a CEO (which might determine the fate of a company; Gomulya et al., 2016; Graham, Harvey, & Puri, 2017; Stoker, Garretsen, & Spreuwiers, 2016), to the sentencing of a criminal (which might determine the fate of a human being; Blair, Judd, & Fallman, 2004; Eberhardt et al., 2006; J. P. Wilson & Rule, 2015). It may thus not be surprising that researchers have called for a more nuanced understanding of why people persistently rely on face judgments (Olivola, Funk, et al., 2014). We contribute to this debate by showing that peoples' reliance on face judgments is better explained by the intuitive accessibility of the cue, which make reliance on it relatively effortless, rather than by beliefs that facial cues are particularly valid. We recommend that future attempts at designing interventions focus on disrupting the primary and efficient processing of faces to attenuate the effects of face judgments on social decisions.



# Chapter 7

Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions

Based on:

Jaeger, B., Todorov, A. T., Evans, A. M., & van Beest, I. (2019). *Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions* Manuscript submitted for publication.

All data, preregistration documents, and analysis scripts are available at the Open Science Framework (<https://osf.io/cbsmw/>).

### **Abstract**

Despite their poor diagnostic value, trait impressions from faces influence many consequential decisions. Here, we test the effectiveness of two types of interventions in reducing the influence of facial stereotypes. We first introduce a legal decision-making paradigm that measures reliance on facial appearance at the participant level. Results of a pretest (Study 7.1,  $n = 320$ ) show that defendants with an untrustworthy (vs. trustworthy) facial appearances are found guilty more often. We then test the effectiveness of the different interventions in reducing the influence of facial stereotypes. Educating participants about the biasing effects of facial stereotypes reduces the explicit belief that personality is reflected in facial features, but it does not reduce the influence of facial appearance on verdicts (Study 7.2,  $n = 979$ ). In Study 7.3 ( $n = 975$ ), we present information sequentially to disrupt the intuitive accessibility of trait impressions. Participants indicate an initial verdict based on case-relevant information and a final verdict based on all information (including facial photographs). This intervention actually *increases* the influence of facial appearance on verdicts. Together, our findings highlight the persistent influence of facial appearance on decision-making.

People spontaneously judge a person's character based on their facial appearance and these judgments guide many consequential decisions (Olivola, Funk, et al., 2014). For instance, trustworthiness impressions affect legal decision-making (J. P. Wilson & Rule, 2015), personnel selection (Gomulya et al., 2017), and economic exchanges involving trust (Rezlescu et al., 2012). People even rely on facial stereotypes when other, more diagnostic cues are available (Jaeger, Evans, Stel, et al., 2019a; Olivola et al., 2018). The widespread influence of trustworthiness impressions is somewhat surprising, given that the accuracy of trustworthiness judgments based on facial features is weak at best (Bonnefon et al., 2017; Todorov, Funk, et al., 2015; Todorov, Olivola, et al., 2015).

Similar patterns emerge for other personality traits. For example, competent-looking people are favored as business leaders, even though they do not seem to perform better (Graham et al., 2017; Stoker et al., 2016). Thus, people persistently rely on facial stereotypes even though these stereotypes are generally inaccurate. This overreliance on facial appearance can lead to worse decision outcomes and systematic discrimination against people with a certain appearance (Olivola & Todorov, 2010b). As a consequence, researchers have called for efforts to mitigate this bias (Olivola, Funk, et al., 2014; Porter et al., 2010; J. P. Wilson & Rule, 2015). Here, we answer this call by exploring the effectiveness of two different types of interventions in reducing reliance on facial stereotypes.

### **Facial stereotypes influence decision-making**

While there are numerous studies demonstrating the effects of facial stereotypes, comparatively little is known about *why* people persistently rely on trait impressions. Recently, two (non-mutually exclusive) hypotheses have been put forward to address this gap. One explanation posits that the widespread influence of trait impression can

be explained by lay beliefs in the diagnostic value of facial appearance for inferring personality traits (Jaeger, Evans, Stel, et al., 2019b; Rezlescu et al., 2012; Todorov, 2017). Crucially, such beliefs may drive reliance on facial stereotypes because how much people rely on a certain cue is usually not determined by how predictive the cue actually is (i.e., how accurate trait impressions are), but by how predictive people *think* the cue is (i.e., how accurate people think their trait impressions are, Brunswik, 1956; Hammond, Hursch, & Todd, 1964). Recent investigations have shown that physiognomic beliefs (i.e., beliefs are indeed common among laypeople (Jaeger, Evans, Stel, et al., 2019b; Suzuki et al., 2017). Moreover, individual differences in physiognomic beliefs predict reliance on trait impressions when making economic trust decisions (Jaeger, Evans, Stel, et al., 2019b): People who more strongly believe that trustworthiness is reflected in facial features rely more on their counterpart's perceived trustworthiness when deciding whom to trust. Thus, reliance on facial stereotypes may be driven by beliefs in the diagnostic value of facial appearance for judging a personality.

A second explanation posits that the intuitive accessibility of trait impressions can account for their persistent effects (Jaeger, Evans, Stel, et al., 2019a). Faces attract attention (Ro et al., 2001; Theeuwes & Van der Stigchel, 2006) and are processed quickly and efficiently (Stewart et al., 2012; Willis & Todorov, 2006). This processing advantage leads to an intuitive accessibility of trait impressions. In line with this reasoning, reliance on facial stereotypes is relatively fast and not influenced by the restriction of cognitive capacities (Bonnefon et al., 2013; Jaeger, Evans, Stel, et al., 2019a; Mieth et al., 2016). Crucially, previous research has shown that people favor readily available cues as they reduce decision effort (Evans & Krueger, 2016; Gigerenzer et al., 2011; Shah, 2007; Shah & Oppenheimer, 2008). Thus, people may rely on trait impressions because this allows them to make decisions relatively effortlessly.



## Reducing reliance on facial stereotypes

To sum up, we posit that the pervasive influence of facial stereotypes may be driven by a combination of (a) false beliefs about the diagnostic value of facial appearances and (b) the intuitive accessibility of trait impressions. Crucially, similar mechanisms have been identified in related research areas that investigate sources of biased behavior. Theories in the field of judgment and decision-making often distinguish between two general sources of bias: false beliefs (i.e., misconceptions) and automatically activated associations (i.e., misleading intuitions; Morewedge & Kahneman, 2017; Soll, Milkman, & Payne, 2014; T. D. Wilson & Brekke, 1994). Moreover, social psychological theories of bias typically distinguish between explicit and implicit expressions of bias (Devine, 1989; Dovidio, Kawakami, & Gaertner, 2002; Greenwald & Banaji, 1995; Greenwald, McGhee, Jordan, & Schwartz, 1998). Due to these similarities, we draw on the extensive literature on debiasing techniques in judgment and decision-making (Morewedge et al., 2015; Soll et al., 2014) and social psychology (Forscher et al., 2019; Lai et al., 2014) to design interventions aimed at reducing reliance on facial stereotypes.

A prominent strategy for addressing bias stemming from misconceptions is to educate people about their false beliefs (Chan, Jones, Hall Jamieson, & Albarracín, 2017; Soll et al., 2014). For example, educating people about compound interest can increase saving behavior (McKenzie & Liersch, 2011), educating people about cognitive biases can lead to more rational clinical decision-making (Hershberger, Markert, Part, Cohen, & Finger, 1997), and raising awareness of prejudice based on social group affiliation can reduce discrimination (Axt, Casola, & Nosek, 2018). Relatedly, confronting participants with their stereotypes, rather than just raising awareness about the existence of stereotypes in general, can reduce bias (Czopp, Monteith, & Mark, 2006; Parker, Monteith, Moss-Racusin, & Van Camp, 2018). In Study 7.2, we therefore

tested whether we can reduce reliance on facial stereotypes by educating people about the influence of facial stereotypes or by confronting them with the fact that their facial stereotypes are not accurate.

A prominent strategy for reducing bias caused by automatically activated associations is to design the decision environment in such a way that participants are nudged to rely on the “right” information (Soll et al., 2014; Thaler & Sunstein, 2008). The primary and efficient processing of faces leads to a quick availability of face-based inferences (Freeman & Johnson, 2016; Todorov et al., 2009; Willis & Todorov, 2006). Crucially, information that is available first often exerts a disproportionate influence on decisions (Asch, 1946; Dimov & Link, 2017; Sullivan, 2018). Initial response tendencies that are based on intuitively accessible cues are often not sufficiently adjusted based on subsequently processed information (producing anchoring effects; Tamir & Mitchell, 2012; Tversky & Kahneman, 1974). Moreover, people are sometimes not able or willing to exert the cognitive effort required to integrate all available information (Shah & Oppenheimer, 2008; Simon, 1955). As a consequence, they make decisions based on the cue that was processed first in order to reduce decision effort (Gigerenzer et al., 2008).

Thus, trait impressions from faces may be influential because people can rely on them to make quick and effortless decisions. This implies that manipulating how deeply and in which order information is processed could reduce the influence of facial stereotypes. In Study 7.3, we therefore tested whether preventing the primary processing of faces by presenting information sequentially (with faces being displayed after more relevant information) reduces reliance on facial stereotypes. We also tested whether prompting participants to make reflective rather than intuitive decisions can reduce bias.

We are not the first to test how different factors influence reliance on facial stereotypes. Providing information on how trustworthy a

person has been in the past (Rezlescu et al., 2012) or giving feedback about a person's trustworthiness in a repeated interaction (Yu et al., 2014) can reduce reliance on facial trustworthiness. In a similar vein, simply omitting photos from the decision-making environment would obviously eliminate the influence of facial appearance. These strategies may be effective, but they are not viable interventions in most real-world situations. When deciding on the culpability of a defendant or on the suitability of a job candidate, decision-makers are often faced with a limited amount of ambiguous or contradicting pieces of information, and it may not be possible to provide additional information about past behavior.

It might also not be possible to completely remove information about a person's appearance from the situation. For these reasons, and in contrast to previous work, we focused on interventions that do not omit or add any additional decision-relevant information. In other words, we aimed at testing the effectiveness of different interventions under conditions that resemble the real-world situations in which the biasing effect of facial appearance is particularly prevalent and problematic (e.g., in criminal sentencing, personnel selection, voting).

### **The current studies**

We present the results of three studies. In a pretest (Study 7.1,  $n = 320$ ), we develop and validate a legal sentencing paradigm that measures reliance on facial appearance. We focus on decision-making in a legal context, because sentencing decisions can be immensely consequential, making biased decision-making particularly problematic (Berry & Zebrowitz-McArthur, 1988; Eberhardt et al., 2006; Porter et al., 2010; J. P. Wilson & Rule, 2015). We then test the effectiveness of four interventions in reducing reliance on facial appearance in two preregistered studies. In Study 7.2 ( $n = 979$ ), we test two informational interventions to reduce the explicit belief that facial appearance is an

indicator of personality. In Study 7.3 ( $n = 975$ ), we test two interventions to disrupt the intuitive accessibility of trait impressions by presenting information sequentially and by inducing a reflective decision-making style. All data, materials, preregistrations, and analysis scripts are available at the Open Science Framework (<https://osf.io/h4yf3/>). We report how our sample sizes were determined, all data exclusions, and all measures in the studies.

### **Study 7.1: Pretest**

Our first goal was to create a legal sentencing task that allows us to measure reliance on facial appearance. Previous experimental studies have predominantly taken two methodological approaches when assessing the influence of facial appearance on verdicts. In some studies, participants view a series of face images and indicate sentencing decisions (e.g., Wilson & Rule, 2016). Multiple trials with within-subjects manipulations of facial appearance increase statistical power, but providing little or no other background information on the cases limits the ecological validity of the task. In other studies, participants receive realistic case descriptions including relevant extenuating or aggravating facts (e.g., Berry & Zebrowitz-McArthur, 1988; Gunnell & Ceci, 2010). This approach more closely resembles the conditions in which decisions are made in real life. However, these studies usually focus on between-subject designs with one or a few cases and face images, limiting statistical power and the generalizability of the results.

Here, we tried to balance advantages of the two approaches. Based on descriptions of real small claims court cases, we created ten fictitious case files, with plaintiffs filing suits against defendants. Cases included realistic evidence and we manipulated the perceived trustworthiness of plaintiffs and defendants in a within-subjects design. Participants indicated sentencing decisions for all ten cases. In line with previous studies, we expected participants to find defendants guilty more often

when they look untrustworthy (vs. trustworthy). We also measured confidence in verdicts and, in case participants ruled in favor of the plaintiff, how much damages they wished to award.

## Methods

**Participants.** We recruited a total of 363 US American workers from Amazon Mechanical Turk (MTurk; Paolacci & Chandler, 2014) who participated in exchange for \$1.50. Data from 30 participants (8.26%) who failed an attention check at the end of the study and 8 participants (2.40%) who indicated having only a poor or basic English proficiency, leaving a final sample of 325 participants (50.46% female,  $M_{\text{age}} = 35.91$ ,  $SD_{\text{age}} = 10.03$ ).

**Materials.** We created case files for ten fictitious small claims court cases (see Figure 7.1). Case files included a photo and demographic information for the plaintiff and the defendant. All individuals were White, male US citizens and had their first and last name redacted. Case files also included the size of the plaintiff's claim (ranging from \$600 to \$3,600) and a case summary of approximately 130 words. Each summary mentioned the reason why the plaintiff was suing the defendant (e.g., seeking reimbursement for a damaged stereo system) and the evidence that was presented by the plaintiff and the defendant (e.g., photos of a broken speaker, a receipt confirming the purchase of a stereo system). In line with real-world small claims court cases, the evidence presented by both sides was relatively limited and weak.

We selected 20 images of White male individuals from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015). The database includes ratings of all targets on various trait dimensions. Based on these data, we selected the ten individuals who received the lowest ( $M = 2.62$ ,  $SD = 0.17$ ) and highest ( $M = 3.78$ ,  $SD = 0.09$ ) ratings on perceived trustworthiness. The targets varied in perceived age with average age ratings ranging from 19.5 to 43.2 years ( $M = 28.60$ ,  $SD = 6.90$ ). We used Psychomorph

(Tiddeman, Burt, & Perrett, 2001) to increase or decrease perceived trustworthiness. For each trustworthy (untrustworthy) target, we transformed the face shape towards a computer-generated trustworthy (untrustworthy) face prototype by 60% (Oosterhof & Todorov, 2008). This procedure somewhat exaggerated the facial features linked to perceptions of trustworthiness and allowed us to create prototypically (un-)trustworthy-looking individuals without compromising the realistic nature of the face stimuli.

Finally, we matched case files and face images. Each case featured a plaintiff and defendant that differed on perceived trustworthiness: One individual looked trustworthy while the other looked untrustworthy. We created four sets of stimuli. Each set contained all ten case files and all 20 face images. In each set, face images were randomly matched to a case and a role (i.e., plaintiff or defendant). Half of all cases featured a trustworthy-looking plaintiff and an untrustworthy-looking defendant, while the roles were reversed in the other half.



	
<b>Plaintiff</b>	<b>Defendant</b>
First name: [REDACTED] Last name: [REDACTED]	First name: [REDACTED] Last name: [REDACTED]
Sex: Male Citizen: U.S. Race: Caucasian	Sex: Male Citizen: U.S. Race: Caucasian
<b>Claim: \$1,700</b>	
<b>Case summary:</b>	
<p>The plaintiff is seeking reimbursement for damages made to a stereo system that was lent to the defendant for a party. The plaintiff presented a receipt for the purchase of a stereo system worth \$1,400 from four months ago. The plaintiff also presented photos showing a broken speaker and a video showing that the system is not functioning anymore. Photos of the system suggested that fluids were spilled on it. The defendant stated that the plaintiff knew that some damages might have occurred during the party. The defendant claimed that the plaintiff had agreed to watch over the system during the party to make sure that nothing happened to it and that the plaintiff is suing him because he needs the money to pay off a debt. The plaintiff denied this and stated that it was the defendant's responsibility to ensure that no damage was caused during the party.</p>	

Figure 7.1. A case file with a trustworthy-looking plaintiff and an untrustworthy-looking defendant.

**Procedure.** Participants were randomly assigned to one of the four stimulus sets. To measure sentencing decisions, participants were told to read each case carefully and to recommend a sentence by ruling in favor of the plaintiff or the defendant. After each ruling, participants also indicated their confidence in the ruling on a scale that ranged from 1 (*not confident at all*) to 9 (*extremely confident*). In case participants ruled in favor of the plaintiff, they were asked to indicate the amount of damages that the plaintiff should be awarded on a scale that ranged from 50% to 100% (in steps of 10%) of the original claim.

**Sensitivity analysis.** We conducted a post hoc sensitivity analysis to determine the smallest effect size we were able to detect for our main effect of interest (the effect of facial trustworthiness on verdicts) with 80% power (and  $\alpha = 5\%$ ). As software commonly used for sensitivity

analyses, such as G\*Power (Faul et al., 2007), does not support multilevel data, we relied on the *simr* package (Green & Macleod, 2016) in R (R Core Team, 2019). The package provides power estimates for fixed effects in multilevel regression models. We systematically varied the effect of facial trustworthiness on verdicts and calculated power at each level, to test which effect size we were able to detect with at least 80% power. This showed that we had 80% power to detect an odds ratio of 1.27 for the effect of facial trustworthiness on verdict. To illustrate, an odds ratio of this size corresponds to a six percentage point difference in guilty verdicts (e.g., 50% vs. 56%) for trustworthy-looking versus untrustworthy-looking defendants.

## Results

On average, participants found the defendant guilty 53.26% of the time ( $SD = 18.54\%$ ). Prevalence of guilty verdicts also varied across cases ( $SD = 13.12\%$ ), with no case receiving a guilty verdict less than 25% or more than 75% of the time.

We analyzed the effect of facial trustworthiness on sentencing decisions by estimating a multilevel regression model with random intercepts and slopes per participant and per case. This accounts for variation in the overall rate of guilty verdicts across participants (i.e., some participants indicating more guilty verdicts than others) and across cases (i.e., some cases receiving more guilty verdicts than others). Regressing verdict (0 = defendant is not guilty, 1 = defendant is guilty) on facial trustworthiness (-0.5 = trustworthy-looking defendant, 0.5 = untrustworthy-looking defendant) revealed a positive effect,  $\beta = 0.319$ ,  $SE = 0.080$ ,  $z = 3.94$ ,  $p < .001$ , 95% CI [0.161, 0.477],  $OR = 1.38$ . The rate of guilty verdicts was 8.03 percentage points higher for untrustworthy-looking defendants (56.65% vs. 48.61%).

We also explored whether facial appearance affected confidence in verdicts or the amount of money participants awarded to the plaintiff in



case of a guilty verdict. Regressing confidence on facial trustworthiness, verdict, and their interaction showed a positive effect of a guilty verdict,  $\beta = 0.243$ ,  $SE = 0.052$ ,  $t(3,016) = 4.75$ ,  $p < .001$ , 95% CI [0.142, 0.344]. Participants were more confident in their verdicts when they ruled in favor of the plaintiff. There was no effect of facial trustworthiness,  $\beta = -0.020$ ,  $SE = 0.071$ ,  $t(1,099) = 0.29$ ,  $p = .77$ , 95% CI [-0.159, 0.118], and no interaction effect between verdict and facial trustworthiness,  $\beta = 0.088$ ,  $SE = 0.099$ ,  $t(2,990) = 0.89$ ,  $p = .37$ , 95% CI [-0.106, 0.281].

Finally, regressing the amount of money that was awarded to the plaintiff in case of a guilty verdict on facial trustworthiness revealed a positive effect,  $\beta = 1.655$ ,  $SE = 0.727$ ,  $t(137.4) = 2.28$ ,  $p = .024$ , 95% CI [0.191, 3.078]. Participants awarded the plaintiff 2.02% more of their original claim when the defendant looked untrustworthy.

## Discussion

Results showed that legal sentencing decisions were influenced by the facial trustworthiness of the involved parties. The rate of guilty verdicts was 7.90 percentage points higher when the defendant looked untrustworthy (vs. trustworthy). Facial trustworthiness also influenced how much money participants awarded to the plaintiff in case of a guilty verdict, with plaintiffs receiving 2.00% more when they were suing an untrustworthy-looking (vs. trustworthy-looking) defendant. We did not find any evidence that the confidence in verdicts was influenced by facial appearance. In sum, using a novel sentencing task with multiple cases and controlled manipulations of facial trustworthiness, we replicate prior work showing that people rely on facial appearance to make legal sentencing decisions (Porter et al., 2010; J. P. Wilson & Rule, 2015; Zebrowitz & McDonald, 1991).

### **Study 7.2: Belief interventions**

In Study 7.2, we used the same sentencing task to test the effectiveness of two interventions in reducing reliance on facial trustworthiness. To achieve this, we aimed to reduce explicit beliefs that personality can be judged from facial appearance (Jaeger, Evans, Stel, et al., 2019b). In one condition, participants read a text that informed them about scientific research on facial stereotypes. The text mentioned the automatic accessibility of facial stereotypes, how facial stereotypes are usually not accurate, and how they nonetheless affect decision-making. The intervention specifically focused on facial stereotypes, as previous work suggests that raising awareness of stereotypes in general may not be effective (Axt et al., 2018). Our manipulation was modelled after previous research in the domain of lay beliefs. For instance, Levy and colleagues (1998) used fake scientific articles to manipulate beliefs about the innateness of personality traits and this influenced how strongly participants associated different social groups with stereotypical personality traits.

In a second intervention condition, we additionally confronted participants with the low diagnostic value of their facial stereotypes. Before reading the educational text, we showed participants ten pairs of faces. Their task was to identify which of the two individuals was a convicted felon. We told participants that they only guessed four out of ten correctly, meaning that their guesses were not better than chance. We measured physiognomic beliefs (i.e., participants' explicit beliefs that personality traits can be judged accurately from faces) in all conditions and hypothesized that both interventions would reduce physiognomic beliefs and reliance on facial trustworthiness when making sentencing decisions.

## Methods

**Power analysis.** We conducted an a priori power analysis using the *simr* package in R, which allows one to test how power varies as a function of the number of random effects levels (in our case, the number of participants or the number of cases). As the number of cases was fixed, we tested how power varies across different numbers of participants. Calculating power across a wide range of sample sizes showed that 250 participants per condition are required to detect a 30% decrease in the effect of facial trustworthiness on verdicts with 80% power (and  $\alpha = 5\%$ ). As a conservative measure, we decided to recruit 325 participants per condition.

**Participants.** We recruited a total of 1,249 US American workers from Amazon Mechanical Turk who participated in exchange for \$2.50. Data from 227 participants (18.17%) who failed an attention check at the end of the study and from 42 participants (4.11%) who indicated poor or basic English proficiency were excluded from analysis, leaving a final sample of 979 participants (47.40% female,  $M_{\text{age}} = 36.14$ ,  $SD_{\text{age}} = 11.24$ ).

**Materials and procedure.** Participants were randomly allocated to one of three conditions. In all conditions, participants completed the legal sentencing task as described in the previous study. For each case, they ruled in favor of the plaintiff or the defendant and their confidence in the ruling on a scale that ranged from 1 (*not confident at all*) to 9 (*extremely confident*). Next, to measure belief in the visibility of personality traits in facial appearance, participants completed the physiognomic belief scale (Jaeger, Evans, Stel, et al., 2019b). Participants were prompted to imagine seeing the passport photo of a stranger. They were asked to indicate how much they agree with three statements (e.g., *I can learn something about a person's personality just from looking at his or her face*) on a scale from 1 (strongly disagree) to 7 (strongly agree).

Average scores across the three items constituted our measure of physiognomic beliefs (Cronbach's  $\alpha = .84$ ).

The three conditions only differed in the texts participants were exposed to prior to completing the sentencing task. In the education condition ( $n = 332$ ), participants read an educational text about personality impressions from faces that was approximately 300 words long. Among other things, the text mentioned that people spontaneously form impressions of others based on their facial appearance, that these impressions contain little accuracy, and that they nonetheless influence many important decisions (the exact text can be found in the online materials).

In the education-and-confrontation condition ( $n = 332$ ), prior to reading the educational text, participants completed an additional task that was designed to demonstrate that their face-based impressions could be inaccurate. Participants saw ten pairs of faces of male individuals. The images were taken from the 10k Faces Database (Bainbridge, Isola, Blank, & Oliva, 2013). Participants were told that each pair included one convicted felon and that their task was to identify that person. Feedback about accuracy was standardized across all participants. They were told that they only guessed four out of ten correctly, meaning that their guesses were not better than chance.

In the control condition ( $n = 315$ ), participants read a text about the geography of Scotland.

After reading the respective texts, participants answered three comprehension check questions (e.g., *Research shows that first impressions influence many important decisions*). Participants could only proceed to the sentencing task after having answered all three questions correctly.

## Results

Participants found the defendant guilty 51.47% of the time ( $SD = 17.47\%$ ). Three participants (0.31%) found all defendants guilty, whereas four (0.41%) found none guilty. Prevalence of guilty verdicts also varied across cases ( $SD = 10.40\%$ ), with no case receiving a guilty verdict less than 25% or more than 75% of the time.

**Physiognomic beliefs.** First, we tested whether the interventions reduced beliefs that personality is reflected in facial appearance. Compared to participants in the control condition ( $M = 3.80, SD = 1.37$ ), participants in the education condition ( $M = 3.59, SD = 1.33$ ) indicated lower physiognomic beliefs,  $t(640.6) = 2.03, p = .042, d = 0.16$ , and so did participants in the education-and-confrontation condition ( $M = 3.50, SD = 1.38$ ),  $t(643.5) = 2.80, p = .005, d = 0.22$ . These results show that both interventions were somewhat successful in reducing belief that personality is reflected in facial features, although differences were small. Physiognomic beliefs did not significantly differ between the education and education-and-confrontation condition,  $t(661.2) = 0.82, p = .41, d = 0.06$ .

**Sentencing decisions.** Next, we tested whether the interventions reduced reliance on facial trustworthiness in the legal sentencing task. We estimated a multilevel regression model with random intercepts and slopes per participant and case. Regressing verdict (0 = defendant is not guilty, 1 = defendant is guilty) on facial trustworthiness (-0.5 = trustworthy-looking defendant, 0.5 = untrustworthy-looking defendant), condition (with the control condition being the reference category), and their interaction terms revealed a positive effect of facial trustworthiness,  $\beta = 0.327, SE = 0.101, z = 3.22, p = .001, 95\% CI [0.127, 0.527], OR = 1.39$ . There were no significant differences in guilty verdicts between the control condition and the education condition,  $\beta = 0.107, SE = 0.061, z = 1.74, p = .083, 95\% CI [-0.014, 0.227], OR = 1.11$ , or the

education-and-confrontation condition,  $\beta = 0.059$ ,  $SE = 0.061$ ,  $z = 0.96$ ,  $p = .34$ , 95% CI [-0.062, 0.179],  $OR = 1.06$ .

Crucially, examining the interaction effects showed that the effect of facial trustworthiness was not significantly different in the education condition,  $\beta = 0.132$ ,  $SE = 0.133$ ,  $z = 0.99$ ,  $p = .32$ , 95% CI [0.130, 0.395],  $OR = 1.14$ , or in the education-and-confrontation condition,  $\beta = -0.056$ ,  $SE = 0.134$ ,  $z = 0.42$ ,  $p = .68$ , 95% CI [-0.318, 0.207],  $OR = 0.95$  (see Figure 7.2). The difference between the education condition and the education-and-confrontation condition was also not significant,  $\beta = -0.188$ ,  $SE = 0.132$ ,  $z = 1.42$ ,  $p = .16$ , 95% CI [-0.448, 0.072],  $OR = 0.83$ . Thus, neither intervention was successful in reducing reliance on facial trustworthiness. The rate of guilty verdicts was 7.78 percentage points higher for untrustworthy-looking defendants in the control condition (54.76% vs. 46.98%), 11.34 percentage points higher in the education condition (58.71% vs. 47.37%), and 6.65 percentage points higher in the education-and-confrontation condition (54.69% vs. 48.04%).

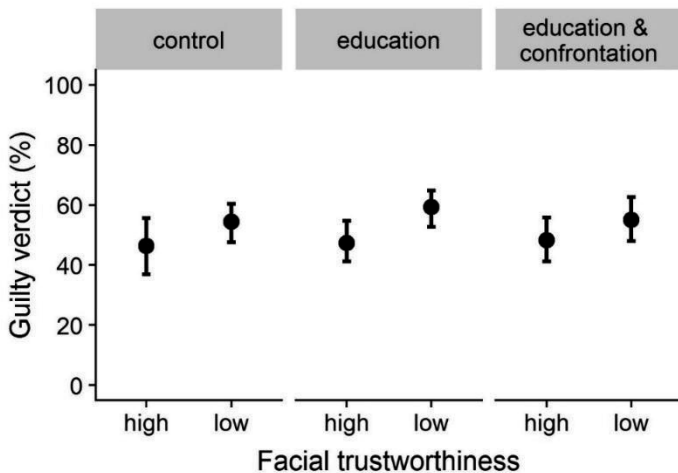


Figure 7.2. Differences in rates of guilty verdicts for trustworthy-looking and untrustworthy-looking defendants as a function of condition. Dots denote predicted values. Error bars denote bootstrapped 95% confidence intervals.

**Confidence in verdicts.** We also tested whether the interventions influenced confidence in verdicts. Regressing confidence on facial trustworthiness, verdict, and condition yielded a positive effect of a guilty verdict,  $\beta = 0.180$ ,  $SE = 0.030$ ,  $t(9,076) = 5.99$ ,  $p < .001$ , 95% CI [0.120, 0.238]. As in Study 7.1, participants were more confident in their verdicts when they found the defendant guilty. There was no effect of facial trustworthiness,  $\beta = 0.033$ ,  $SE = 0.049$ ,  $t(6.25) = 0.68$ ,  $p = .52$ , 95% CI [-0.070, 0.136], and compared to the control condition, confidence was not significantly different in the education condition,  $\beta = -0.001$ ,  $SE = 0.091$ ,  $t(973.0) = 0.01$ ,  $p = .99$ , 95% CI [-0.180, 0.178], or in the education-and-confrontation condition,  $\beta = -0.145$ ,  $SE = 0.091$ ,  $t(973.4) = 1.59$ ,  $p = .11$ , 95% CI [-0.323, 0.034]. In other words, we did not find evidence that the interventions influenced confidence in verdicts.

**Exploratory analyses.** The design of the interventions tested here was based on a proposed link between belief in the visibility of personality in a person's facial features and reliance on facial appearance when making decisions (Jaeger, Evans, Stel, et al., 2019b). Even though the interventions somewhat reduced physiognomic beliefs, they did not reduce reliance on facial trustworthiness, raising the question whether physiognomic beliefs were related to reliance on facial trustworthiness in the current study. To test this, we extracted participant-specific slopes for the effect of facial trustworthiness from our multilevel regression models, as an indicator of how much each participant relied on facial appearance when making sentencing decisions. There was indeed a significant correlation between physiognomic beliefs and reliance on facial trustworthiness,  $r(977) = .200$ ,  $p < .001$ . There was also a positive correlation between physiognomic beliefs and confidence in verdicts,  $r(977) = .204$ ,  $p < .001$ . Participants who believed more strongly that personality is reflected in facial features relied more on facial trustworthiness when making sentencing decisions and they were more confident in their verdicts.

To further probe the effects of the two interventions, we conducted Bayesian analyses using the *BayesFactor* package (Morey & Rouder, 2018) in R (R Core Team, 2019). Bayesian *t*-tests with default Cauchy priors yielded substantial support for the null hypothesis of no difference between the control condition and the education condition,  $BF_{01} = 6.49$ , and strong support for the null hypothesis of no difference between the control condition and the education-and-confrontation condition,  $BF_{01} = 10.66$ . These results support the conclusion that neither intervention significantly reduced reliance on facial appearance.

## **Discussion**

Results showed that neither intervention successfully reduced reliance on facial stereotypes. Educating people about the low accuracy of their trait impressions did reduce their explicit beliefs in the diagnosticity of facial appearance for judging personality, but this effect was relatively small. Importantly, it did not reduce reliance on facial stereotypes when making sentencing decisions and it also did not reduce their confidence in verdicts. The same pattern was observed for a second intervention in which participants were additionally confronted with the low accuracy of their own trait impressions.

### **Study 7.3: Accessibility interventions**

In Study 7.3, we tested the effectiveness of two alternative interventions in reducing reliance on facial trustworthiness. Trait impressions from faces are intuitively accessible (Stewart et al., 2012; Todorov et al., 2009; Willis & Todorov, 2006) and accessible information often exerts a disproportionate influence on decisions (Shah, 2007; Simmons & Nelson, 2006; Tversky & Kahneman, 1974). To disrupt the primary processing of faces, we presented information sequentially. First, participants saw only case-relevant information and indicated an initial verdict. Then, they saw the entire case file (which also included



face images of the plaintiff and defendant) and indicated their final verdict. We hypothesized that the majority of participants would not revise their initial verdicts. Reliance on intuitively available trait impressions constitutes a low-effort decision strategy and people might not be aware of the extent to which their decisions are influenced by facial stereotypes (Jaeger, Evans, Stel, et al., 2019a). However, in our sequential design, participants have to actively revise their verdict (and ignore case-relevant information) to rely on facial appearance. They might also be reluctant to do so because sticking to their initial (unbiased) verdict should reduce decision effort (Shah & Oppenheimer, 2008; Simon, 1955). Thus, the influence of facial stereotypes should decrease under sequential presentation, because the majority of decisions reflect verdicts that were made when no information on facial appearance was available.

In a second intervention condition, we tested whether the influence of intuitively available trait impressions would be further reduced by prompting participants to make reflective decisions (Newman et al., 2017). To ensure that initial verdicts are based on a careful consideration of the case-relevant information, participants had to reflect on their initial verdicts for a predetermined time.

## Methods

**Participants.** Based on the power analysis reported in Study 7.2, we again decided to recruit 325 participants per condition. We recruited a total of 1,085 US American workers from Amazon Mechanical Turk who participated in exchange for \$2.50. Data from 93 participants (8.57%) who failed an attention check at the end of the study and from 17 participants (1.71%) who indicated poor or basic English proficiency were excluded from analysis, leaving a final sample of 975 participants (49.74% female,  $M_{\text{age}} = 35.86$ ,  $SD_{\text{age}} = 10.50$ ).

**Materials and procedure.** Participants were randomly allocated to one of three conditions. In all conditions, participants again completed the legal sentencing task. For each case, they ruled in favor of the plaintiff or the defendant and indicated their confidence in the ruling on a scale that ranged from 1 (*not confident at all*) to 9 (*extremely confident*).

In the sequential condition ( $n = 319$ ), participants first saw the case files without any personal information about the plaintiff or defendant and were asked to indicate an initial ruling in favor of the plaintiff or the defendant. Next, participants saw the entire case files, including the images of the plaintiff and defendant, and were asked to indicate their final ruling and their confidence in the ruling on a scale that ranged from 1 (*not confident at all*) to 9 (*extremely confident*).

In the sequential-and-reflection condition ( $n = 329$ ), participants followed the same procedure as in the sequential condition, but they were prompted to think carefully and make reflective decisions for all cases (Newman, Gibb, & Thompson, 2017). They could only indicate an initial ruling after 30 seconds had passed and they were instructed to take at least this long to carefully study the case summary before indicating a ruling.

In the control condition ( $n = 327$ ), participants completed the legal sentencing task without the order of stimuli being manipulated.

## Results

Participants found the defendant guilty 52.01% of the time ( $SD = 16.44\%$ ). Five participants (0.51%) found all defendants guilty, whereas four (0.41%) found none guilty. Prevalence of guilty verdicts also varied across cases ( $SD = 13.52\%$ ), with no case receiving a guilty verdict less than 25% or more than 75% of the time.

**Response times.** First, we analyzed response times for initial rulings to check whether instructions to reflect on decision in the sequential-and-reflection condition actually led to longer decisions

times compared to the sequential condition. We excluded 63 response times (0.65%) that were more than three standard deviations above the mean response time from analysis and  $\log_{10}$ -transformed response times due to their right-skewed distribution. A  $t$ -test showed that participants took longer to reach a decision in the sequential-and-reflection condition ( $M = 1.658, SD = 0.111$ ) compared to the sequential condition ( $M = 1.527, SD = 0.273$ ),  $t(417.5) = 7.93, p < .001, d = 0.62$ .

**Sentencing decisions.** Next, we tested whether our interventions reduced reliance on facial trustworthiness in the legal sentencing task. We estimated a multilevel regression model with random intercepts and slopes per participant and case. Regressing verdict (0 = defendant is not guilty, 1 = defendant is guilty) on facial trustworthiness (-0.5 = trustworthy-looking defendant, 0.5 = untrustworthy-looking defendant), condition (with the control condition being the reference category), and their interaction terms revealed a positive effect of facial trustworthiness,  $\beta = 0.218, SE = 0.105, z = 2.08, p = .038, 95\% CI [0.005, 0.431], OR = 1.24$ . There were no significant differences in rates of guilty verdicts between the control condition and the sequential condition,  $\beta = 0.095, SE = 0.058, z = 1.65, p = .10, 95\% CI [-0.018, 0.209], OR = 1.10$ , or the sequential-and-reflection condition,  $\beta = 0.060, SE = 0.057, z = 1.05, p = .30, 95\% CI [-0.053, 0.173], OR = 1.06$ .

Crucially, we found that, compared to the control condition, the effect of facial trustworthiness on verdicts was significantly *larger* in the sequential condition,  $\beta = 0.529, SE = 0.116, z = 4.54, p < .001, 95\% CI [0.301, 0.758], OR = 1.70$ , and in the sequential-and-reflection condition,  $\beta = 0.448, SE = 0.115, z = 3.88, p < .001, 95\% CI [0.222, 0.675], OR = 1.56$  (see Figure 7.3). The difference between the sequential condition and the sequential-and-reflection condition was not significant,  $\beta = -0.081, SE = 0.117, z = 0.69, p = .49, 95\% CI [-0.310, 0.148], OR = 0.92$ . Thus, contrary to our predictions, both interventions significantly *increased* the influence of facial trustworthiness. The rate of guilty verdicts was 5.40

percentage points higher for untrustworthy-looking defendants in the control condition (53.51% vs. 48.11% guilty verdicts), 18.40 percentage points higher in the sequential condition (62.25% vs. 43.85% guilty verdicts), and 16.81 percentage points higher the sequential-and-reflection condition (60.54% vs. 43.73% guilty verdicts).

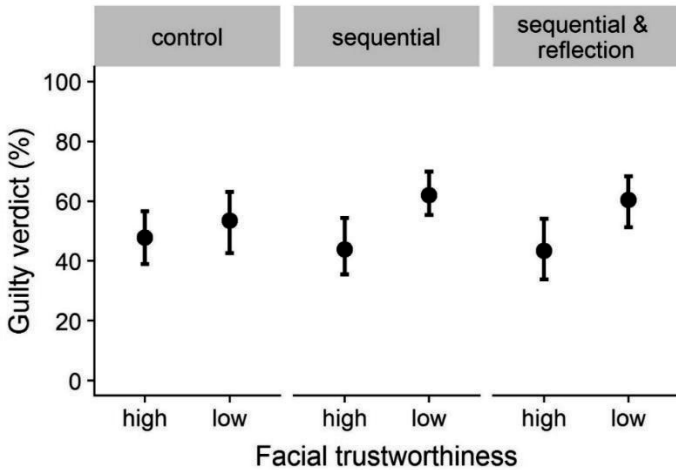


Figure 6.3. Differences in rates of guilty verdicts for trustworthy-looking and untrustworthy-looking defendants as a function of condition. Dots denote predicted values. Error bars denote bootstrapped 95% confidence intervals.

**Confidence in verdicts.** We also tested whether the interventions influenced confidence in verdicts. Regressing confidence on facial trustworthiness, verdict, and condition yielded no effect of facial trustworthiness,  $\beta = 0.042$ ,  $SE = 0.065$ ,  $t(8.58) = 0.65$ ,  $p = .54$ , 95% CI [-0.092, 0.177], but a positive effect of a guilty verdict,  $\beta = 0.086$ ,  $SE = 0.030$ ,  $t(9,023) = 2.81$ ,  $p = .005$ , 95% CI [0.026, 0.145]. Participants were more confident in their verdicts when they found the defendant guilty. Compared to the control condition, confidence was significantly higher in the sequential condition,  $\beta = 0.333$ ,  $SE = 0.093$ ,  $t(971.8) = 3.59$ ,  $p < .001$ , 95% CI [0.151, 0.515], and also in the sequential-and-reflection

condition,  $\beta = 0.402$ ,  $SE = 0.092$ ,  $t(972.3) = 4.73$ ,  $p < .001$ , 95% CI [0.222, 0.583]. There was no significant difference in confidence between the sequential condition and the sequential-and-reflection condition  $\beta = 0.070$ ,  $SE = 0.093$ ,  $t(972.8) = 0.75$ ,  $p = .45$ , 95% CI [-0.112, 0.251]. In other words, the interventions significantly increased confidence in verdicts.

**Exploratory analyses.** To further probe the effects of the two interventions, we again conducted Bayesian analyses. Bayesian  $t$ -tests with default Cauchy priors yielded strong support for the alternative hypothesis that reliance on facial trustworthiness was stronger in the sequential condition compared to the control condition,  $BF_{10} = 1,484$ , and that reliance on facial trustworthiness was stronger in the sequential-and-reflection condition compared to the control condition,  $BF_{10} = 188$ . These results support the conclusion that both interventions significantly increased reliance on facial appearance.

Finally, we analyzed how often and under what conditions participants revised their initial decision to understand why the interventions increased rather than decreased reliance on facial trustworthiness. We hypothesized that most participants would not revise their initial decisions, which were made in the absence of face images and therefore unbiased by facial trustworthiness.

In fact, the majority of initial rulings in the sequential condition (89.78%) and in sequential-and-reflection condition (90.61%) were not revised when participants saw the images of the plaintiff and defendant and had the chance to do so. However, analyzing revision rates showed that participants were more likely to revise their initial ruling when it was not in line with face stereotypes (e.g., a trustworthy-looking defendant being found guilty; 15.4%) than when it was already in line with stereotypes (3.14%),  $\chi^2(1) = 310.2$ ,  $p < .001$ . Of all revised rulings, 83.52% ended up being congruent with face stereotypes whereas only 16.48% were incongruent with face stereotypes. As a consequence, while only 51.11% of all initial rulings made in the absence of face images

were in line with face stereotypes, 57.61% of all final rulings made in the presence of face images were,  $\chi^2(1) = 55.12$ ,  $p < .001$ . In sum, both interventions were successful in producing unbiased rulings in the absence of face images, which were seldom revised when participants did have access the face images. However, the wide majority of revisions that did occur brought decisions in line with face stereotypes. This increased the overall effect of facial appearance on sentencing decisions.

## **Discussion**

Results showed that both interventions increased, rather than decreased, reliance on facial stereotypes. In order to disrupt the primary processing of faces (and the intuitive accessibility of trait impressions), we asked participants to indicate initial decisions that were solely based on relevant information about the cases. They were then shown the entire case files, which also included facial photographs of the plaintiff and defendant, and they could still revise their sentencing decisions. As intended, the majority of participants (ca. 90%) did not change their initial sentences, which means that most final sentences reflected decisions that were made while being ignorant of the plaintiff's and defendant's facial appearance. However, participants who decided to change their initial decisions overwhelmingly did so to bring their final decisions in line with facial stereotypes (e.g., by finding an untrustworthy-looking defendant guilty). The same pattern was observed for a second intervention in which participants were additionally prompted to make reflective decisions.

## **General discussion**

The aim of the current investigation was to test the effectiveness of different interventions in reducing the influence of facial stereotypes on legal decision-making. We created a legal sentencing task in which participants indicated verdicts for multiple small claims court cases and

we manipulated the perceived facial trustworthiness of plaintiffs and defendants. In line with previous studies showing that facial appearance influences legal decision-making (Blair, Judd, & Chapleau, 2012; Eberhardt et al., 2006; Porter et al., 2010; J. P. Wilson & Rule, 2015; Zebrowitz & McDonald, 1991), we found that defendants were more likely to be found guilty when they looked untrustworthy (vs. trustworthy). This effect was observed in all three studies.

We then tested the effectiveness of four different debiasing techniques (education, confrontation, sequential presentation, and reflection induction) in reducing the influence of facial trustworthiness on verdicts. In Study 7.2, we (a) educated participants about the biasing influence of facial stereotypes and (b) confronted them with the low diagnostic value of their own trait impressions. Both education alone and a combination of education and confrontation reduced explicit beliefs that personality traits are reflected in facial appearance (although these effects were comparatively small). However, neither intervention significantly affected how much participants relied on facial stereotypes when making sentencing decisions. Bayesian analyses indicated strong support in favor of null effects. It may simply be the case that our manipulation was not strong enough to reduce behavioral reliance on facial stereotypes. In past studies, relatively short educational texts were sufficient to influence lay beliefs and related behaviors (Chiu et al., 1997; Levy et al., 1998). Nonetheless, our intervention only had a small effect on beliefs and future studies could test whether more intensive debiasing trainings are more effective (Devine, Forscher, & Austin, 2013; Morewedge et al., 2015; Sellier, Scopelliti, & Morewedge, 2019).

In Study 7.3, we attempted to disrupt the intuitive accessibility of trait impressions by providing information sequentially. First, participants saw only case-relevant information and indicated a preliminary sentence. Then, participants saw the entire case file (including facial photographs) and indicated their final sentence. As

intended, only a minority of initial sentences were changed. However, sentence revisions were strongly driven by facial appearance, with most revised decisions reflecting a stereotype-congruent verdict (e.g., untrustworthy-looking defendants being found guilty). On average, this actually *increased* the influence of facial stereotypes. A similar pattern was observed when participants were additionally prompted to make reflective decisions.

Why did a later presentation of photographs increase the influence of facial stereotypes? Studies on the role of fluency in cue ordering (Dimov & Link, 2017), anchoring effects (Tamir & Mitchell, 2012; Tversky & Kahneman, 1974), and primacy effects in impression formation (Asch, 1946) often highlight the influence of information that is processed *first*. However, other investigations found a disproportionate influence of information that is processed *last* (i.e., recency effects; for an overview see Sullivan, 2018). For example, when evaluating faces that display a series of expressions, trait impressions are more strongly influenced by the expression that was displayed last (Fang, van Kleef, & Sauter, 2018). In a similar vein, participants might have attributed more importance to facial photographs because they were the only new information that was displayed after they had indicated their preliminary verdicts. To participants, this may imply that this information is relevant for their decisions (Clark & Haviland, 1977). More research is needed to systematically explore how the order in which facial appearance and other cues are processed affects the influence of facial stereotypes.

Together, our results highlight the persistent influence of facial stereotypes on decision-making. People rely on facial appearance even when other, more diagnostic cues are available (Jaeger, Evans, Stel, et al., 2019a; Olivola et al., 2018; Olivola & Todorov, 2010b) and even when they are explicitly told not to (Blair et al., 2004). Here, we consistently found that the perceived facial trustworthiness of plaintiffs and



defendants influenced sentencing decisions for small claims court cases. We do not doubt that certain manipulations could diminish or eliminate the effect of facial trustworthiness on verdicts. For example, providing unambiguous, outcome-relevant information (e.g., clear evidence that a defendant committed a crime) has been shown to reduce reliance on stereotypes (Dovidio & Gaertner, 2000; Rezlescu et al., 2012). However, such decisive information is often not available in real life. In many situations, such as legal sentencing, personnel selection, or voting, people have to make consequential decisions based on limited, ambiguous, or contradicting information. For these reasons, we tested the effectiveness of different interventions in a decision-making environment with multiple ambiguous cues.

Evidence for the biasing influence of facial appearance is well-documented and researchers have called for attempts to curb this bias (Olivola, Funk, et al., 2014; Porter et al., 2010; J. P. Wilson & Rule, 2015). We took a first step in this direction but, ultimately, we were unsuccessful in reducing the influence of facial stereotypes. Our choice of interventions was based on recent insights into the mechanisms underlying reliance on facial stereotypes (Jaeger, Evans, Stel, et al., 2019a, 2019b). To identify more successful interventions, future investigations could draw on the rich literature on changing implicit and explicit bias (Forscher et al., 2019; Lai et al., 2014). To stimulate research in this area, we have made all materials needed to implement the legal sentencing task that was used here publicly available. This task allows for within-subject manipulations of facial appearance (or of other cues such as race or gender), which is statistically powerful and provides an indicator of reliance on facial stereotypes at the participant level. We hope that our results will motivate others to design and test different interventions.

## **Conclusions**

In sum, all four interventions tested here (education, confrontation, sequential presentation, and reflection induction) were ineffective in reducing the influence of facial stereotypes on sentencing decisions. Participants consistently found untrustworthy-looking defendants guilty more often than trustworthy-looking defendants. Our results underscore the persistent influence of facial appearance on decision-making. They also highlight the need for more research on the mechanisms underlying reliance on facial appearance and potential ways of reducing bias caused by facial stereotypes.

# Chapter 8

General discussion

People spontaneously judge another person's character based on their facial appearance (Todorov, Olivola, et al., 2015). Even though trait impressions from faces are generally inaccurate, they influence a wide range of social decisions (Olivola, Funk, et al., 2014). For example, due to their facial appearance, a person may be passed over for a promotion (Ling et al., 2019) or receive a lower wage at work (Hamermesh & Biddle, 1994); they be found guilty of a crime (J. P. Wilson & Rule, 2015) or receive a harsher punishment for it (Blair et al., 2012). In short, people not only experience unfair treatment because of their gender, race, or sexual orientation, but also because of their facial appearance. In this dissertation, I therefore argued that first impressions should be treated like other social biases.

A social bias perspective focuses the study of first impressions on three central issues: (a) estimating the *prevalence and magnitude* of the bias, (b) explaining its underlying *mechanisms*, and (c) testing *interventions* that aim to mitigate the bias. While an extensive literature has documented the prevalence and magnitude of facial discrimination, little is known about the underlying mechanisms or potential ways to mitigate it. Why do people persistently rely on trait impression, despite their low diagnostic value and despite the fact that this can lead to worse decision outcomes? How can facial discrimination be reduced? The six empirical chapters presented in this dissertations aimed to address these questions.

### **Summary of main findings**

Chapters 2 and 3 provided new insights into the *prevalence and magnitude* of facial discrimination. We replicated and extended two findings from the literature showing that facial appearance influences decision-making in real life. **Chapter 2** examined the influence of facial appearance on voting behavior. Analyzing election data of 150 mayoral candidates from 75 municipalities across Italy, we found that attractive-

looking (but not competent- or trustworthy-looking) candidates received more votes and were more likely to win. We also tested whether the electoral success of trustworthy-looking politicians varied across regions. We reasoned that trustworthy-looking politicians may be particularly successful in regions that are characterized by high levels of political corruption, but found no support for this hypothesis.

**Chapter 3** examined the influence of facial appearance on online consumer behavior. Analyzing a sample of 1,020 apartments on Airbnb, we tested whether people are willing to pay a premium to stay with attractive-looking or trustworthy-looking hosts. A hedonic pricing model (Rosen, 1974) showed that attractive-looking hosts (but not trustworthy-looking hosts) charged higher prices for similar apartments, suggesting that people favor staying with them. Moreover, we found that Black hosts charge lower prices than White hosts and hosts with more pronounced smiles charge higher prices. Together, our results suggest that consumers rely on various facial characteristics when deciding whom to stay with on Airbnb.

Chapters 4-6 investigated the *mechanisms* underlying reliance on first impressions. In **Chapter 4**, we examined whether reliance on first impressions can be explained by beliefs in the diagnostic value of facial appearance for inferring personality traits (i.e., physiognomic beliefs). Across five samples, including a large, representative sample of the Dutch population, we found that physiognomic beliefs are widespread. Crucially individual differences in physiognomic beliefs are associated with various social-cognitive processes and behaviors. People who more strongly believe in physiognomy are more confident in the accuracy of their first impressions and this increased confidence is not due to the fact that their judgments are more accurate. In addition, they rely more on first impressions when making trust decisions, even when they have access to a more valid cue.

In **Chapter 5**, we replicated the finding that endorsement of physiognomic beliefs correlates with confidence in first impressions. Going beyond previous results, we examined the role of physiognomic beliefs in an applied setting. We found that people who more strongly believe in physiognomy view information on a job candidate's facial appearance as more diagnostic for making hiring decisions, and they view reliance on facial appearance when making hiring decisions as more appropriate and more effective.

We also investigated differences in physiognomic beliefs across personality traits (rather than across individuals). We found that sociability is believed to be more reflected in facial features than morality or competence. This pattern of results also emerged for outcomes associated with physiognomic beliefs: (a) people were more confident in the accuracy of their sociability (vs. morality or competence) judgments; (b) information on job candidates' facial appearance was valued more when looking for a sociable (vs. moral or competent) candidate; and (c) reliance on facial appearance to make hiring decisions was seen as more appropriate and more effective when looking for a sociable (vs. moral or competent) candidate.

Together, findings from Chapter 4 and 5 provide novel insights into a potential mechanism underlying the persistent effects of first impressions. Despite the fact that trait impressions from faces are generally inaccurate, subjective beliefs in diagnostic value of facial appearance for judging personality are widespread. Moreover, across a wide range of outcomes, we found that physiognomic beliefs are correlated with an increased reliance on facial cues in judgment and decision-making.

In **Chapter 6**, we examined another mechanisms that could explain the persistent influence of first impressions. Research in judgment and decision-making shows that biases often arise from a tendency to rely on cues that are intuitively accessible (T. D. Wilson & Brekke, 1994). We

therefore tested whether people rely on trait impressions because of their intuitive accessibility (Engell et al., 2007; Willis & Todorov, 2006), which makes reliance on them relatively effortless. When making trust decisions, participants viewed information on their interaction partner's incentive to betray trust (i.e., their temptation) as more diagnostic than information on their facial appearance. However, making trust decisions based on facial appearance was less effortful than making trust decisions based on temptation. Crucially, when both cues are available, we found a slight preference for relying on facial appearance: participants relied less on temptation when they could rely on facial appearance instead; they did not rely less on facial appearance when they could rely on temptation instead. Together, our results suggest that people rely on first impressions even when (objectively and subjectively) more valid cues are available because reliance on facial appearance is less effortful.

In **Chapter 7**, we leveraged these novel insights into the mechanisms underlying reliance on first impressions to design *interventions* aimed at reducing the influence of facial appearance on decision-making. We created a novel decision-making paradigm that allowed us to measure reliance on first impressions in a legal setting. Participants indicated verdicts for multiple small claims court cases and we manipulated the perceived facial trustworthiness of plaintiffs and defendants. Results showed that participants were more likely to find defendants guilty if they looked untrustworthy (vs. trustworthy). Thus, we replicated previous findings showing that facial appearance influences legal sentencing decisions (Berry & Zebrowitz-McArthur, 1988; Porter et al., 2010).

We then tested the effectiveness of two types of interventions. First, we attempted to reduce the influence of facial appearance on verdicts by educating participants about the low diagnostic value of their impressions. Our intervention was successful in reducing physiognomic beliefs, but it did not reduce the effect of facial appearance on verdicts.

Next, we attempted to disrupt the intuitive accessibility of first impressions by displaying information sequentially. Participants saw only case-relevant information and indicated a preliminary sentence. Then, participants saw the entire case file (including facial photographs) and indicated their final sentence. As intended, only a minority of initial sentences (which were unbiased due to the absence of facial photographs) were changed. However, sentence revisions were strongly driven by facial appearance, with most revised decisions reflecting a stereotype-congruent verdict (e.g., untrustworthy-looking defendants being found guilty). On average, this actually increased the effect of facial appearance on verdicts. Thus, both interventions were unsuccessful in reducing the influence of facial appearance on legal sentencing decisions, which again attests to the persistence of facial discrimination.

### **Theoretical implications**

#### **The consequences of first impressions**

Our findings provide new insights into the prevalence, magnitude, and persistence of facial discrimination. In line with previous findings (Olivola, Funk, et al., 2014), results of all six empirical chapters show that trait impressions from faces influence a broad range of decisions, both in the lab and in real life. People rely on trait impressions when making economic trust decisions (Chapters 4 and 6) and legal sentencing decisions (Chapter 7), when deciding whom to vote for in an election (Chapter 2) and whom to stay with on Airbnb (Chapter 3). These effects were often not trivial in size. For instance, a one standard deviation increase in the facial attractiveness of mayoral candidates in the 2016 Italian local elections corresponded to a 3 percentage point increase in vote share and a 1.9 times increase in the odds of victory. Thus, effects of first impressions are consequential, both in terms of their prevalence and magnitude.



Our results also provide insights into the situational factors that influence when people rely on first impressions. Previous research on the precursors of stereotyping has shown that people often rely on generalized beliefs about others (a) when they do not have access to more diagnostic information (Dovidio & Gaertner, 2000) or (b) when they are not motivated enough to consider additional cues (Neuberg & Fiske, 1987).

If we view first impressions as generalized beliefs about others based on their facial appearance (i.e., as facial stereotypes), we might expect that people would only rely on first impressions in situations where no better cues are available. In contrast to this argument, our studies show that people rely on first impressions even if there are other, more diagnostic cues available. We found effects of facial appearance on Airbnb even though the website displays many relevant cues (e.g., review scores, apartment size) in an easily accessible way.

Similarly, we might expect that people only rely on first impressions when decisions are not consequential enough to justify expanding the additional cognitive effort required for considering other cues. However, we found effects of facial appearance on legal sentencing decisions and real-world voting decisions, both of which should motivate people to make unbiased decisions based on relevant information (see also Graham, Harvey, & Puri, 2017; J. P. Wilson & Rule, 2015). Thus, the effects of first impressions are not only widespread and consequential, but also surprisingly persistent. People rely on facial appearance even when they have access to more diagnostic information and even when they should be motivated to make unbiased decisions.

**The role of facial attractiveness.** What is the relationship between facial attractiveness judgments and judgments of other traits, such as trustworthiness, competence, or dominance? On the one hand, there seem to be clear conceptual differences between attractiveness and other traits. For example, it may seem strange to talk about the

accuracy of attractiveness inferences. Even though there is some consensus in who is perceived as attractive (Hehman et al., 2017), attractiveness judgments are thought to reflect an individual's personal taste, making debates about whose attractiveness ratings are most accurate relatively futile (following the economists' maxim of *de gustibus non est disputandum*).

Moreover, judgments of attractiveness and personality traits are often theorized to occupy different positions in the causal chain between facial appearance and behavior (e.g., Verhulst, Lodge, & Lavine, 2010). Work on the so-called attractiveness halo shows that attractive people are perceived to have more positive personality traits (Dion et al., 1972). Building on this observation, many studies demonstrating positive associations between attractiveness and desirable social outcomes cite the attractiveness halo as the most likely underlying explanation: Attractiveness impressions influence decisions *because* they trigger personality impressions (Maestripietri et al., 2017).

Both views may lead researchers to study the influence attractiveness and personality trait impressions in isolation. On the one hand, researchers who focus on the effects of personality trait impressions might not investigate the role of attractiveness because they view attractiveness as just one cue among many on which these impressions are based. On the other hand, researchers who focus on the effects of attractiveness impressions might not investigate the role of personality trait impressions because—even though they might mediate any effect of attractiveness—attractiveness is the more ultimate explanation in their causal model.

However, there are good reasons why this might be misguided. First, the attractiveness halo explanation has only received partial support: Many effects of facial attractiveness on social outcomes are *independent* of more favorable perceptions on relevant personality dimensions (Maestripietri et al., 2017). In fact, in Chapters 2 and 3 we

found that the influence of attractiveness impressions was not due to impressions on personality dimensions that may have reasonably been expected to explain any effect of attractiveness (e.g., the perceived trustworthiness of Airbnb hosts or the perceived competence of political candidates). Thus, every effect of attractiveness is explained by more favorable personality impressions. Similarly, not every effect of personality impressions is ultimately due to attractiveness (Duarte et al., 2012; Ravina, 2008). These results highlight that attractiveness and personality trait impressions can have dissociable effects and that both should be studied together to disentangle their unique effects.

**Beauty premium vs. ugliness penalty.** Why do people favor attractive individuals in so many domains? Based on an extensive review of the literature on attractiveness biases, Maestripieri and colleagues (2017) concluded that attractiveness biases most likely exist because people are attuned to evaluate the mate value of potential partners and these evaluations can spill over into decision-making domains unrelated to mating. Some faces are more attractive than others because they exhibit characteristics that are seen as indicators of good health or other traits that are desirable in potential mates (Jaeger et al., 2018; Pazda et al., 2016; Rhodes, 2006). People's tendency to seek out attractive individuals is so strong, that we even favor them in situations outside the mating domain (e.g., when deciding whom to hire or whom to vote for).

Since the costs of mistakenly interacting with an unhealthy individual are higher than the costs of foregoing an interaction with a healthy individual, this view actually predicts that predicts that people should be particularly motivated to avoid unhealthy (and therefore unattractive) individuals (Schaller & Duncan, 2007; Zebrowitz et al., 2003; Zebrowitz & Rhodes, 2004). Thus, effects of attractiveness may be driven by negative reactions towards particularly unattractive people, rather than positive reactions towards particularly attractive people.

The distinction between a beauty premium and an ugliness penalty is often not made in the literature and any effect of attractiveness is typically referred to as a beauty premium. In Chapter 3 we tested for both effects and, in line with the theory, we found support for an ugliness penalty, but not for a beauty premium on Airbnb. Previous work on disparities in wages as a function of attractiveness has yielded mixed results (Hamermesh, 2011; Hamermesh & Biddle, 1994). Again, these findings highlight the need for more rigorous tests of theories that aim to explain the underlying causes of attractiveness biases.

### **The mechanisms underlying reliance on first impressions**

Little is known about why people rely on first impressions. Addressing this question was the primary goal of this dissertation. Our results suggest two explanations. First, results of 8 studies ( $N = 5,299$ ; Chapters 4 and 5) with participants from the Netherlands, the United States, and the United Kingdom show that lay beliefs in diagnostic value of facial appearance (i.e., physiognomic beliefs; Aristotle, 1936; Lavater, 1775) are widespread. Around half of all participants at least somewhat endorsed the belief that personality traits are reflected in facial features. Moreover, we found that physiognomic beliefs are associated with increased confidence in the accuracy of first impressions and increased reliance on first impressions in social decision-making. Thus, one explanation for the persistent influence of first impressions is the widespread (erroneous) belief that an individual's facial appearance is a valuable cue for inferring their personality.

Our findings also point to a second explanation: Reliance on first impressions is easy. Evaluating and integrating different pieces of information requires cognitive effort (Bettman et al., 1990), which people are generally averse to expand (Kool, McGuire, Rosen, & Botvinick, 2010; Shah & Oppenheimer, 2008). One way to reduce decision effort is to rely on cues that are easily accessible (Evans &

Krueger, 2016; Shah, 2007; Simmons & Nelson, 2006). First impressions may constitute such as cue because faces are processed quickly and effortlessly (Engell et al., 2007; Willis & Todorov, 2006). In line with this reasoning, that relying on facial trustworthiness is relatively effortless. Stronger reliance on facial appearance when making trust decisions was not associated with longer decision times. In contrast, an increased reliance on an interaction partner's temptation to betray trust led to longer decision times. Moreover, restricting cognitive reflection had no effect on how much participants relied on facial appearance, whereas it reduced (in fact, completely eliminated) reliance on temptation. These results suggest that at least compared to reliance on temptation, reliance on trustworthiness impressions from faces is relatively effortless. In other words, reliance on first impressions can be seen as a heuristic—a mental shortcut that reduces decision effort (cf. Simon, 1955).

These two explanations for facial discrimination should not be considered mutually exclusive. In fact, our studies suggest that the two paths are interrelated and, at least partly, a result of the automatic way in which trait impressions from faces are formed. The automatic accessibility of first impressions may lead people to make quick and unreflective decisions. In line with the idea that people are often motivated to reduce decision effort by relying on cues that are easily accessible (Evans & Krueger, 2016; Shah, 2007), we found that people rely on facial appearance even when they could rely on another, more valid cue instead. Thus, people may make decisions based on intuitively accessible trait impressions and disregard other cues altogether, even if these cues are more diagnostic.

However, even when people take the time to carefully consider and weigh all available cues before making a reflective decision, they may still end up relying on facial appearance. This is due to the fact that the fluency with which cues are processed influences their perceived diagnosticity (Hertwig, Herzog, Schooler, & Reimer, 2008; Shah, 2007;

Simmons & Nelson, 2006). Put differently, people are confident in the accuracy of their trait impressions *because* they are intuitively accessible. In line with this reasoning, we found that beliefs in the diagnostic value of trait impressions were more pronounced among people who tend to trust their intuition. Thus, the intuitive accessibility of trait impressions promotes their influence on decisions in a direct way—because people may ignore other cues and rely on easily accessible trait impressions to make quick and unreflective decisions—and in an indirect way—because easily accessible trait impressions are perceived as more diagnostic, they are weighed more heavily when making reflective decisions.

**Common causes of decision biases.** The two proposed explanations for reliance on first impressions converge with existing theories on the underlying causes of decision biases. Specifically, research in the area of judgment and decision-making shows that biases often arise due to false beliefs (i.e., misconceptions) or due to automatically activated associations (i.e., misleading intuitions; Morewedge & Kahneman, 2017; Soll, Milkman, & Payne, 2014; T. D. Wilson & Brekke, 1994). Social psychological research on stereotypes and discrimination often distinguishes between biases that results from relatively conscious and reflective decision processes (i.e., explicit bias) and biases that result from relatively unconscious and intuitive decision processes (i.e., implicit bias; Devine, 1989; Dovidio, Kawakami, & Gaertner, 2002; Greenwald & Banaji, 1995; Greenwald, McGhee, Jordan, & Schwartz, 1998).

In a similar vein, we find evidence for two processes that contribute to the persistent reliance on first impressions. On the one hand, people put too much weight on first impressions when making decisions due to false beliefs about their diagnostic value (the reflective route). On the other hand, people rely on first impressions because of their intuitive accessibility, which allows them to make decisions relatively quickly and

effortlessly (the intuitive route). In other words, similar to other types of biased decision-making, facial discrimination can result from both reflective and intuitive decision processes.

**Bottom-up and top-down influences in impression formation.**

Our results also converge with recent findings highlighting the importance of lay beliefs in social cognition (Stolier, Hehman, & Freeman, 2020). To understand impression formation from faces, previous work has predominantly focused on how variations in different facial features, such as babyfacedness, facial width-to-height ratio, or resemblances to emotional expressions, affect trait impressions (Hehman et al., 2019). However, trait impressions are not only influenced by characteristics of the perceived face (i.e., bottom-up, stimulus-driven processes), but also by characteristics of the perceiver (i.e., top-down, perceiver-driven processes; Hehman et al., 2017). For example, the extent to which a perceiver believes that two personality traits are correlated (i.e., whether sociable people also tend to be moral) influences the extent to which their impressions of the two traits are based on the same set of facial features (i.e., whether their sociability impressions are correlated with their morality impression Stolier, Hehman, Keller, et al., 2018). Put differently, the same facial appearance can trigger different trait impressions depending on the perceiver's lay personality theory.

In a similar vein, we find evidence that both reliance on first impressions can be explained by characteristics of the perceived stimulus (i.e., the fluency with faces are processed; Chapter 6) and by characteristics of the perceiver (i.e., lay beliefs in physiognomy; Chapters 4 and 5). These findings are in line with the view that bottom-up, stimulus-driven processes and top-down, perceiver-driven processes both play an important part in impressions formation (Freeman & Ambady, 2011).

## **Practical implications**

The results presented in this dissertation highlight the social costs of first impressions. In many everyday situations, people are being treated differently because of their facial appearance. In this sense, first impressions operate akin to other social biases: People may experience unfair treatment because of their gender, race, or sexual orientation, but also because of their facial appearance. For example, in Chapter 2, we found a price penalty not only for Black Airbnb hosts, but also for those with an unattractive facial appearance (Chapter 3). In a similar vein, people may receive unfair treatment in the legal system because of their race (Steffensmeier, Ulmer, & Kramer, 1998), but also because they are perceived as untrustworthy (Chapter 7; J. P. Wilson & Rule, 2015).

Biased decision-making not only leads to unfair treatment, but also to worse outcomes for decision-makers. People persistently rely on first impressions when deciding whom to trust, even if they have access to a cue that actually predicts the likelihood that their trust will be reciprocated (Chapter 6). This shows that (dis-)trust in others is often misplaced because people erroneously perceive an interaction partner as untrustworthy due to their facial appearance. That is, people are more likely to experience betrayal or miss out on potentially advantageous interactions because they rely on first impressions when deciding whom to trust. In a similar vein, previous studies have shown that people make sub-optimal hiring decisions because some candidates are erroneously perceived as more competent than others (Ling et al., 2019; Stoker et al., 2016). In short, reliance on first impressions leads to unfair treatment and sub-optimal decision-making.

Recognizing first impressions as social biases implies that we should strive to mitigate the negative consequences of facial discrimination. In situations in which these biases often manifest (e.g., in personnel selection), decision-makers need to ensure a fair and unbiased decision-making process by safeguarding against the influence of first



impressions. For this to happen, people need to be aware of how the potential bias resulting from exposure to a person's facial appearance (Axt et al., 2018). This step is especially important because the available evidence suggests that people are often not aware of how first impressions can influence their behavior (Hassin & Trope, 2000). Gender and race constitute salient social categories and gender and racial biases are widely discussed, making it more likely that people will recognize when their behavior may be influenced by these factors.

Moreover, people are very accurate in identifying a person's gender or race (Bruce & Young, 2012), which suggests that it is relatively easy for people to identify when they are interacting with a person that might trigger biased behavior. In contrast, the influence of first impressions is more subtle due to their perceptually ambiguous nature. For instance, a series of studies by Blair and colleagues (2002; 2004) showed that stereotypes can be triggered by categories and features: People were more likely to attribute aggressive behaviors to African Americans (vs. European Americans), but also to individuals with more Afrocentric facial features (i.e., facial features that are typical for African Americans). Crucially, explicitly instructing participants to avoid racial stereotypes reduced the effect of race category, but not the effect of race features. Thus, discrimination based on facial features may be more difficult to identify than discrimination based on salient social categories.

### **Mitigating the consequences of first impressions**

Going beyond awareness of the problem, how can people ensure that decisions are unbiased by facial appearance? When possible, decision-makers should blind themselves to the facial appearance of others. There are also many situations in which information on a person's facial appearance is (a) not relevant for the decision process and (b) easy to conceal. For example, when an HR manager evaluates the résumés of different job candidates, when a judge reads the description

of a case for the first time, or when an Airbnb host decides whether or not to accept a reservation, they should be blind to the appearance of the people they are evaluating. This is already common practice for other types of information. Many academic journals conceal the names and affiliations of authors during the review process. In a similar vein, employers are discouraged to ask questions about a candidate's political or sexual orientation. These mechanisms are in place to prevent that knowledge about certain characteristics, consciously or unconsciously, leads to biased decision-making. A similar reasoning should apply to information on a person's facial appearance. Ultimately, this means that facial photographs should be eliminated from résumés, legal case files, Airbnb profiles, and other types of descriptions.

However, there are several reasons why eliminating information on facial appearance may not be effective or feasible in many real-world situations. First, eliminating face-to-face contact or personal photos can come with costs, which have to be weighed against the potential costs of facial discrimination. For example, granting people the right to represent themselves in court necessarily involves face-to-face contact with those who might end up making biased decisions as a consequence of their first impressions. The prominent display of profile photos on Airbnb provides another good example of this trade-off. An increasing number of studies, including evidence presented in this dissertation, shows that the presence of personal photos leads to discrimination against hosts based on their race or attractiveness (Chapter 3; Edelman & Luca, 2014; Kakar, Franco, Voelz, & Wu, 2016; Wang, Xi, & Gilheany, 2015). While some have urged Airbnb to address the issue by eliminating photos (Edelman, 2016; Edelman & Luca, 2014), Airbnb has argued that personal photos are important for establishing trust between hosts and potential guests (Murphy, 2016).<sup>31</sup> Thus, in many situations, people may be reluctant to

---

<sup>31</sup> Ultimately, the company decided to retain profile photos but to decrease their prominence in order to focus consumers on more valid information (Edelman,

eliminate information about a person's facial appearance due to the associated costs.

Second, even when structural changes to the decision-making environment are made to avoid facial discrimination, people may actively seek out information about a person's facial appearance, and this information is often readily available. A recent field experiment found that attractive applicants received more callbacks irrespective of whether a photo was included in their résumé, or whether it was omitted from their résumé but could be found online (Baert, 2018). Given that people often share facial photographs on Facebook, Twitter, LinkedIn, and other social networking platforms, it is easy for decision-makers to find out what a person looks like with the use of a few pieces of personal information such as their name and place of residence.

Third, face-to-face encounters are often incidental and impossible to avoid. While there is some form of control over what information is displayed in certain decision-making environments (e.g., on Airbnb or when evaluating résumés of job candidates), this control is largely absent in many everyday situations. People interact with strangers on a daily basis and each face-to-face encounter can potentially lead to biased behavior.

Given that eliminating information of a person's facial appearance is often not possible or too costly, alternative strategies for reducing reliance on first impressions are necessary. In Chapter 7, we took a first step in this direction by testing the effectiveness of two types of interventions in reducing the effects of first impressions on decision-making. Even though the interventions were ultimately unsuccessful, they provide an example of how the issue can be approached. Specifically, interventions can focus on changing the decision-maker or the decision-making environment (Soll et al., 2014). Chapters 4 and 5

---

2016; Murphy, 2016); it remains to be seen whether this measure is effective in curbing photo-based discrimination.

suggest that first impressions are influential because people believe that a person's facial features are indicative of their personality. Therefore, interventions that educate people about the low diagnostic value of first impressions may prove effective. On the other hand, results of Chapter 6 suggest that first impressions are influential because they are intuitively accessible and allow people to make decisions quickly and effortlessly. Therefore, interventions that disrupt the accessibility of first impressions (e.g., by nudging people to process other information first) may prove effective. In general, researchers designing intervention studies need to be mindful of the specific mechanism underlying the bias in order to successfully combat it (Lai et al., 2014).

### **Limitations and open questions**

The work conducted as part of this dissertation was motivated by the observation that in many ways, first impressions operate similar to other social biases. For example, people think that men are better scientists than women and, even though this stereotype might not be accurate, it leads them to discriminate against women when making hiring decisions (Régner, Thinus-Blanc, Netter, Schmader, & Huguet, 2019). In a similar vein, people associate certain facial features with personality traits and, even though these impressions are generally inaccurate, they rely on their impressions when making decisions (Graham et al., 2017; Ling et al., 2019). Viewing first impressions as social biases raises novel research questions which this dissertation sought to address. Even though the present results afford a better understanding of various issues, such as why people persistently rely on first impressions, many open questions remain.

### **When do people rely on first impressions?**

Reliance on first impressions extends beyond different cultures (Rule et al., 2010) and age groups (Charlesworth et al., 2019; Suzuki,

2016). People even rely on first impressions when making extremely consequential decisions (J. P. Wilson & Rule, 2015) and when other (more diagnostic) information is available (Olivola & Todorov, 2010b). In line with these previous studies, our results show that the effects of first impressions are widespread, persistent, and often not trivial in size.

Whereas evidence for the claim that people often rely on first impressions is strong, more research is needed to explore *which* trait impressions people rely on in specific situations. Following the seminal work by Todorov and colleagues (2005), dozens of studies have demonstrated that the facial appearance of politicians predicts their electoral success. While most studies found that competent-looking politicians are more successful, some studies also found relationships between electoral success and perceived attractiveness (Berggren, Jordahl, & Poutvaara, 2010; Chapter 2), dominance (F. F. Chen et al., 2014; Sussman et al., 2013), or trustworthiness (F. F. Chen et al., 2016; Rule et al., 2010). Similarly inconsistent results between specific trait impressions and associated outcomes have been observed in financial decision-making (Duarte et al., 2012; Ravina, 2008) and consumer behavior (Chapter 3; Ert, Fleischer, & Magen, 2016).

Which trait impressions do people rely on in specific situations? This should be determined by which trait is most relevant for the specific decision at hand. For instance, a person who is concerned for their safety might prefer a trustworthy-looking Airbnb host, whereas a person who is looking for a good time might prefer an attractive-looking host. More work is needed to determine which contextual or individual differences moderate reliance on specific trait impressions.

An exploration of moderating factors may also help explain which people rely on trait impressions (and not on other cues). We found that people do not rely less on first impressions when provided with a more diagnostic cue (Chapter 6), which is in line with previous investigations showing that the effects of first impressions persist even when people

have access to more valid information (Olivola et al., 2018; Olivola & Todorov, 2010b). However, other work has shown that reliance on facial appearance can be reduced or eliminated by providing feedback about a person's actual behavior in a repeated interaction (Chang, Doll, van 't Wout, Frank, & Sanfey, 2010; Yu et al., 2014) or by providing information about a person's past behavior (Rezlescu, Duchaine, Olivola, & Chater, 2012; but see Li, Liu, Pan, & Zhou, 2017). It remains unclear how abundant, accessible, or diagnostic alternative information needs to be to prevent people from relying on first impressions. Thus, more studies in which different aspects of the decision-making environment are manipulated are needed to explore the boundary conditions of facial discrimination.

One limitation of the current set of studies is that they focused on personality impressions that were formed based on facial photographs. Studying behavior in the presence of personal photos is interesting, as photos are common in many real-life settings ranging from the legal system (e.g., mug shots in case files or offender databases), to the work environment (e.g., company websites, résumés, or online platforms like LinkedIn), to a variety of services that aim to connect people (e.g., Facebook, Tinder, or Airbnb). Yet, there are also many situations in which people encounter strangers and see more than their face. Does facial appearance affect personality impressions and, in turn, decision-making, even when people have access to other cues that can be appraised at a glance or after a very brief encounter such as voice, clothing, or bodily appearance? Evidence from studies examining impression formation based on multiple cues suggests that this is the case. Faces explain a substantial amount of variance in overall judgments and personality judgments based on faces, voices, and bodies are often correlated (M. Peters et al., 2007; Rezlescu et al., 2015; Tsankova et al., 2015). Thus, a focus on faces is justified by their central role in impression formation. Nonetheless, more studies involving actual face-

to-face encounters are needed to understand how people form first impressions based on a wide variety of cues that may hold information about an individual's personality (for an example, see Satchell, 2019).

### **Why do people rely on first impressions?**

Identifying the specific conditions that moderate the strength of facial discrimination will also provide more insights into why people persistently rely on first impressions in spite of their generally low accuracy. Across different samples and outcomes, we found that lay beliefs in the diagnostic value of facial appearance are widespread and correlated with an increased weighing of facial cues in judgment and decision-making (Chapters 4 and 5). In other words, first impressions are more consequential when decision-makers believe that facial features are indicative of personality. In Chapter 5, we found that participants who score higher on physiognomic belief rely more strongly on facial appearance when making trust decisions even when another, more diagnostic cue was present. In Chapter 7, we conceptually replicated this relationship: belief in physiognomy was positively correlated with reliance on facial appearance in a legal sentencing task. In both studies, the observed effect size was relatively small ( $r \approx .20$ ) and future studies are needed to probe the robustness of this result. Moreover, results of Chapter 7 showed that educating people about the low diagnostic value of facial appearance reduced physiognomic beliefs, but did not reduce reliance on first impressions. The relatively small decrease in physiognomic beliefs suggests that the manipulation might not have been strong enough to change people's behavior. Additional work is needed to establish a causal relationship between physiognomic beliefs and reliance on first impressions.

Modest associations between physiognomic beliefs and reliance on facial cues suggest that there are other factors that lead people to rely on first impressions. In fact, results of Chapter 6 suggest that people rely on

first impressions because of the efficient way in which faces are processed, which allows them to make quick and effortless decisions. However, it is still unclear how the intuitive accessibility of trait impressions contributes to the influence of facial appearance in decision-making. People may minimize decision effort by terminating cue search early (Shah & Oppenheimer, 2008). That is, they may make a decision based on trait impressions because they come to mind first and ignore other cues altogether.

Alternatively, people may consider all available information, but the primary processing of faces may change how subsequently processed information is interpreted or weighed. For example, people may seek out information that confirms their first impressions or discount disconfirming evidence (Nickerson, 1998). Future studies could employ process tracing methods such as Mouselab to investigate which cues people focus on in what order and how this influences the weighing of different cues (Johnson, Payne, Schkade, & Bettman, 1989).

### **How can we reduce reliance on first impressions?**

Increased knowledge about the mechanisms underlying facial discrimination will also help in designing more effective interventions. Eliminating information about a person's facial appearance is probably the most effective strategy to curb facial discrimination (but see Baert, 2018). In a similar vein, providing people with diagnostic information that is displayed in an accessible way may also be effective (Rezlescu et al., 2012; Yu et al., 2014). However, these measures are often impossible to implement, which means that alternative strategies that prevent people from relying on first impressions are required.

While the interventions that were tested in Chapter 7 did not succeed in reducing reliance on first impressions, they provide a useful blueprint for future investigations. First, the legal sentencing paradigm that we used provides a convenient testbed for studying the impact of



different interventions. It enables researchers to measure reliance on first impressions at the participant level with a task that includes more naturalistic stimuli than previously used paradigms (such as the trust game Rezlescu et al., 2012; Yu et al., 2014) in a domain in which facial discrimination is particularly problematic (Blair et al., 2012; Eberhardt et al., 2006; J. P. Wilson & Rule, 2015).

Second, the interventions that were tested in Chapter 7 were designed to address the specific mechanisms that have been shown to underlie reliance on first impressions. While none of the interventions were successful, certain design changes may improve their efficacy. For instance, our results suggest that reducing beliefs in the diagnostic value of facial appearance should lead less reliance on first impression (Chapter 4). Thus, a more intensive intervention which demonstrates to participants that their first impressions are largely inaccurate (but nonetheless influence their behavior) could be more successful.

### **Future directions for the study of first impressions**

The study of first impressions can be broadly divided into three research areas: (a) the formation of first impressions (Which traits do people infer from faces? Which cues do they use to make these inferences?), (b) the accuracy of first impressions (How accurate are trait inferences from faces? What can explain this accuracy?), and (c) the consequences of first impressions (When do first impressions influence behavior? Why do people rely on first impressions?). The present dissertation focused on the third research area, as relatively little was known about why people rely on first impressions when making social decisions. Given the widespread negative consequences of facial discrimination, this questions is of particular importance and a first step towards curbing this bias. Nevertheless, there are many important questions that remain unanswered in the other two research areas.

### **How do people form trait impressions from faces?**

Many studies have been devoted to identifying the specific facial characteristics that make a person look trustworthy, dominant, or attractive (Todorov, Olivola, et al., 2015). Researchers have focused on a wide array of cues, including morphological (e.g., facial width-to-height ratio; Stirrat & Perrett, 2010), statistical (e.g., gender- or race-typicality; Blair et al., 2002; Walker et al., 2017), and textural (e.g., skin smoothness; Jaeger, Wagemans, Evans, & van Beest, 2018), and demographic features (e.g., gender and race; Xie, Flake, & Hehman, 2018). However, recent investigations have shown that trait impressions are not only determined by which face is perceived, but also by who is perceiving it (Hehman et al., 2017). Understanding who is more likely to perceive others as trustworthy, dominant, or attractive is crucial for predicting trait impressions, but little research has been conducted on this topic thus far.

One promising line of inquiry is to test whether the functional significance of trait impressions—which have mostly been tested by measuring or manipulating aspects of the perceived stimulus (Radke, Kalt, Wagels, & Derntl, 2018)—is also evident when considering aspects of the perceiver. For instance, trustworthiness impressions are thought to constitute an evaluation of a person's intentions, which then motivate basic approach-avoidance tendencies (Todorov, Said, Engell, & Oosterhof, 2008). At the perceiver level, this view would predict that individual differences in approach-avoidance motivations (cf. Carver & White, 1994; Elliot & Thrash, 2002) should be associated with trustworthiness impressions from faces. That is, perceivers who score high on approach-related traits should perceive the same face as more trustworthy than perceivers who score low on approach-related traits. In general, future studies need to go beyond examining characteristics of the face and focus on characteristics of the perceiver (Sprenghelmeyer et

al., 2016) or the situation (Brambilla et al., 2018) to explain how trait impressions from faces are formed.

### **How accurate are trait impressions from faces?**

Even though many studies have investigated this question, evidence on whether traits such as trustworthiness (Bonnefon et al., 2013; Efferson & Vogt, 2013), extraversion (Ames et al., 2010; Borkenau et al., 2009), or competence (Graham et al., 2017; Rule & Ambady, 2008) can be inferred from facial features is inconclusive. Future studies would benefit from addressing key methodological limitations that plague many existing studies.

First, studies should employ cropped face images that were taken under standardized conditions to rule out that judgment accuracy is driven by non-face cues, such as clothing (X. Wu & Zhang, 2016) or hairstyle (W. T. L. Cox et al., 2015). More ecologically valid approaches that mimic the conditions in which people make personality judgments are undoubtedly useful, as people rarely encounter disembodied faces in real life. For instance, Van de Ven and colleagues (2017) examined accuracy in personality judgments based on LinkedIn profiles, which not only includes personal photos but also a variety of other cues. Nonetheless, more controlled studies are necessary to support the claim that facial features in particular provide information about an individual's personality.

Second, many existing studies relied on relatively small samples of face images, which limits their statistical power and generalizability. Accuracy in personality judgments generally increases with the richness of the available information and cropped images of still faces constitute a relatively poor stimulus (Borkenau & Liebler, 1992; Funder, 2012). Future studies should therefore include large number of targets and raters to ensure sufficient statistical power to detect small effects. This would also increase the generalizability of results, which is particularly

important given that several studies that found accuracy in trustworthiness detection relied on the same set of face images, meaning that these results cannot be treated as independent evidence (Bonnefon et al., 2013; Centorrino, Djemai, Hopfensitz, Milinski, & Seabright, 2015; De Neys, Hopfensitz, & Bonnefon, 2013; De Neys et al., 2017).

Finally, many existing studies found accuracy in judgments only under very specific conditions (and these conditions differed across studies). For example, Tognetti and colleagues (2013) found that observers were able to predict the cooperativeness of male, but not female targets. In a similar vein, Verplaetse and colleagues (2007) found that raters accurately identified cooperative partners from facial photographs that were taken at the moment when partners were deciding whether or not to cooperate, but not from photos that were taken during a practice round or before the start of the study. These results might reveal important boundary conditions for judgment accuracy. However, they may also represent false positives, as increasing the number of tests that are conducted to examine a specific hypothesis (by splitting the data into different subgroups) is bound to produce a statistically significant effect (Simmons et al., 2011). Researchers can avoid this problem by specifying any factor that is expected to moderate the effect of interest a priori (i.e., by preregistering their studies; Nosek, Ebersole, DeHaven, & Mellor, 2018). This should not discourage researchers to conduct exploratory analyses, especially if their studies are sufficiently powered to detect the presence of potential moderators (Frankenhuis & Nettle, 2018). However, if exploratory analyses reveal an effect that is deemed interesting, confirmatory studies should then be conducted to ensure the replicability of the result.

In sum, future studies should employ large samples of diverse faces that were photographed under standardized conditions, which are then evaluated by a large samples of raters. Moreover, preregistration and replication can enhance the credibility of results, which will aid in

building a strong empirical foundation. Recent studies in the field of social perception have already started adopting these standards and can serve as a blueprint for future work (e.g., Jones et al., 2019; Lin, Adolphs, & Alvarez, 2018).

## Conclusion

The work that was conducted as part of this dissertation suggests four general conclusions. First, *in spite of their generally low accuracy, trait impressions based on facial features influence a wide range of social decisions*. Effects of first impressions on decision-making are prevalent, often not trivial in size, and surprisingly persistent. These findings imply that people are often treated unfairly because of their facial appearance. In this sense, the effects of first impressions are comparable to other forms of social biases, such as discrimination based on a person's gender, race, or sexual orientation.

Second, *people persistently rely on first impressions because they believe in the diagnostic value of facial appearance for judging personality*. Many people believe that a person's facial appearance is indicative of various personality traits. Moreover, these beliefs are associated with an increased reliance on first impressions in judgment and decision-making.

Third, *people persistently rely on first impressions because it allows them to make decisions quickly and effortlessly*. People are often motivated to avoid cognitive effort. One way to reduce cognitive effort when making decisions is to rely on cues that are intuitively available and, because faces are processed quickly and efficiently, trait impressions from faces constitute such an accessible cue. Thus first impressions can be viewed as a heuristic—a decision strategy that allows people to save cognitive resources.

Fourth, insights into the mechanisms underlying reliance on first impressions will help future research to identify interventions that

mitigate facial discrimination. Even though the specific interventions that were tested in this dissertation were ultimately unsuccessful, our results suggest that *reliance on facial appearance may be reduced by educating people about the low diagnostic value of their first impressions or by disrupting the intuitive accessibility of first impressions*. Together, these findings advance our understanding when and why people rely on first impressions to make decisions. They also have important practical implications, as any attempt at mitigating facial discrimination will require an understanding of the underlying mechanisms.

The two surrealist paintings by René Magritte that are displayed on the cover of this book (in slightly adapted versions) elegantly capture the main themes of this dissertation. The speed and ease with which we can judge people's character just from looking at their faces is remarkable. Due to the intuitive availability of first impressions, it is very tempting to let them guide our behavior. This (almost) irresistible influence of first impressions is symbolized by the painting on the front cover (*The Son of Man*, 1964). However, looks can be deceiving. Our evolved psychology predisposes us to appraise other people's intentions and abilities based on whatever information we have at our disposal. This sensitivity to pick up on signals from faces can misfire and lead us to see signals where there are none. This idea is captured by the painting on the back cover (*The Telescope*, 1963). Generally speaking, we should try to resist the allure of first impressions. Just like we should not judge a person by the color of their skin, we should not judge them by the morphology of their face.

# Appendix

## Automated classification of demographics from face images: A tutorial and validation

Based on:

Jaeger, B., Slegers, W. W. A., & Evans, A. M. (in press). Automated classification of demographics from face images: A tutorial and validation. *Social and Personality Psychology Compass*.

All data and analysis scripts are available at Open Science Framework (<https://osf.io/23pn4>).

## Abstract

Examining disparities in social outcomes as a function of gender, age, or race has a long tradition in psychology and related disciplines. With an increasing availability of large naturalistic data sets, researchers are afforded the opportunity to study the effects of demographic characteristics with real-world data and high statistical power. However, since demographic characteristics are often determined by having participants rate images of targets, limits in participant pools can hinder researchers from analyzing large data sets. Here, we present a tutorial on how to use two face classification algorithms, Face++ and Kairos. We also test and compare their accuracy under varying conditions and provide practical recommendations for their use. Drawing on three face databases ( $n = 3,141$  images), we find that classification accuracy of the algorithms is (a) generally high and similar to the accuracy of human raters, (b) similar for standardized and more variable images, and (c) dependent on various factors such as the target's race, the angle from which targets were photographed, and which algorithm is used. In sum, we propose that automated face classification can be a useful tool for researchers interested in studying the effects of demographic characteristics in large naturalistic data sets.



Across the social sciences, researchers are interested in how demographic characteristics shape social outcomes. People spontaneously encode a person's gender, age, and race (Bruce & Young, 2012; Fiske, Haslam, & Fiske, 1991) which, in turn, triggers a wealth of stereotypes that can influence judgments and decisions (Eagly & Wood, 1999; Fiske, Cuddy, Glick, & Xu, 2002). In fact, gender-, age-, or race-based disparities have been observed in a wide variety of domains such as legal decision-making (C. S. Jones & Kaplan, 2003; Petsko & Bodenhausen, 2019; Steffensmeier et al., 1998), economic exchange (Ayres & Siegelman, 1995; Belot, Bhaskar, & van den Ven, 2012; Eckel & Grossman, 1998), and the work environment (Gordon & Arvey, 2004; Rupp, Vodanovich, & Credé, 2006). In short, exploring the systematic differences in how people behave or are treated by others as a function of their gender, age, or race has a long tradition in psychology, as well as in related fields such as economics, sociology, and law.

To study the effects of demographic characteristics, researchers often draw on large naturalistic data sets. For example, scholars have investigated data from game shows (Belot, Bhaskar, & van de Ven, 2010; Darai & Grätz, 2013), dating websites (Feliciano, Robnett, & Komaie, 2009), criminal trials (Blair et al., 2012; Starr, 2014; Steffensmeier et al., 1998), and online peer-to-peer markets (Doleac & Stein, 2013; Edelman et al., 2017). These investigations are part of the emerging field of computational social science, which uses big data to answer questions relevant to social scientists (Lazer et al., 2009).

Relying on large naturalistic data sets has several advantages. It allows for a more precise estimation of effect sizes and provides a more direct test of how demographic variables influence outcomes in real-life settings. While creating such data sets can be very time-intensive, researchers can often draw on existing data sets that were created for purposes other than psychological research. In addition, more researchers are sharing their data thanks to recent initiatives promoting

open science, enabling others to test novel hypotheses using openly available data sets (Kidwell et al., 2016).

Despite the increasing availability of large data sets, researchers often have to focus on subsets of the available data due to resource constraints (e.g., Kakar et al., 2016). Since information on targets' demographic characteristics is often not available, researchers typically rely on human raters to code demographic information based on face images. This is a valid approach, as people are able to identify a person's gender, age, and race with very high levels of accuracy (Bruce & Young, 2012). However, the required sample of raters vastly outnumbers the typical university participant pool. For example, acquiring ratings for 100,000 images by 15 independent judges on three characteristics requires a participant pool of 22,500 individuals.<sup>32</sup> It would be difficult to reach this sample size, even with access to online participant pools, such as Amazon Mechanical Turk (MTurk; Paolacci & Chandler, 2014). Here, we propose that one solution to this problem is to rely on automated procedures that classify a target's gender, age, and race from face images.

### **Automated face classification**

While automated face classification has received considerable attention in the computer science literature (Gutta, Huang, Jonathon, & Wechsler, 2000; Levi & Hassner, 2015; Lu & Jain, 2004), social scientists have only recently begun to incorporate the technology into their research (e.g., An & Weber, 2016; Edelman, Luca, & Svirsky, 2017; Huang, Weber, & Vieweg, 2014; Jadidi, Karimi, Lietz, & Wagner, 2017; Kosinski, 2017; Messias, Vikatos, & Benevenuto, 2017; Rhue & Clark, 2016). For example, Edelman and colleagues (2017) found that hosts on Airbnb were less likely to accept requests by potential guests with

---

<sup>32</sup> This calculation assumes that each participant takes 20 minutes to rate a total of 200 images on one characteristic.

stereotypically Black names. They used Face++, a face classification algorithm, to determine the race of hosts' previous guests and showed that the pro-White bias in acceptance rates was lower for those who had already hosted a Black guest in the past. In a similar vein, Rhue and Clark (2016) compiled data on more than 100,000 entries on Kickstarter—a crowdfunding platform where people can raise capital for various projects. Face++ was used to identify the demographic characteristics of fundraisers, with the finding that projects of Black fundraisers attracted fewer contributions and were less likely to succeed in meeting their required funding.

Relying on an algorithm instead of participants offers several key advantages. It allows researchers to work with large data sets and reduces the time spent on data collection. We therefore posit that automated face classification can be a useful tool for researchers interested in studying the effects of demographic characteristics on social outcomes. The remainder of this article is organized in three parts. First, we provide a short tutorial on how to use face classification algorithms. A more detailed tutorial including annotated R code is provided in the Supplemental Materials.

Second, in two studies, we assess and compare the accuracy of two algorithms in categorizing gender, age, and race based on face images. In a first study, we draw on two face databases with 597 and 2,208 targets to test the algorithms' accuracy for standardized and more variable images. In a second study, we draw on a third database with 336 targets to test whether the algorithms' accuracy is affected by two attributes of the photographed targets: head orientation and facial expression. We also recruit human raters to provide age, gender, and race classifications, allowing us to directly compare the algorithms' performance against the performance of human raters.

Third, we discuss several advantages and disadvantages of relying on face classification algorithms, and provide practical recommendations for researchers interested in using the algorithms.

### **How to use face classification APIs**

Here, we focus on two face classification algorithms: Face++ (Megvii Inc., <http://www.faceplusplus.com>) and Kairos (Kairos AR, Inc., <https://www.kairos.com>). We are in no way affiliated with the companies providing the services. While there might be a number of different algorithms, we focus on Face++ and Kairos for three reasons: Both can be easily accessed via the openly available software R (R Core Team, 2019); they can classify—among other things—a target’s gender, age, and race; and they have a variety of pricing plans.

Face++ and Kairos can be accessed via their respective Application Programming Interface (API). An API is a way of accessing the functionality of a program via another program. APIs usually have their own website where users can access their functionality (for examples, see the demo pages of both Face++ and Kairos). Another way of accessing the functionality of an API is via code. Rather than manually clicking and uploading photos, users can instruct a computer program to perform this task, resulting in an “API call”. An API call consists of a communication between a client (i.e., a user’s computer) and a server (i.e., the place where the API-related computations are performed).

To enable this communication, many APIs rely on the Hypertext Transfer Protocol (HTTP), which can perform two basic actions: GET and POST. GET methods involve a request for data, while POST methods involve sending data to the server to be processed, after which data is returned. In the case of both Face++ and Kairos, POST methods are used since the purpose is to send a specific photo to their servers and retrieve the photo’s attributes. The returned data is formatted in the JSON format. JSON is a text format that structures the data in a way that is easy to read for programs such as R (but less easy to read for humans). Additional

steps should then be taken to parse the data and prepare it for analysis. In short, an API call consists of sending a message to a server (e.g., sending an image to Face++ or Kairos), which performs the requested computations (e.g., classifying demographic characteristics), and returns the results back to the user.

Because API calls are requested computations, there is often a set of controls in place that prevent the API from being overused or abused. One such control is the use of API keys. Users have to create an account and obtain a set of API keys in order to use the API. API keys are used to track and control how and by whom the API is used. Often, there is a public key (similar to a username) and a secret key (similar to a password). Once the keys are obtained, the API can be accessed with a variety of popular programming languages (e.g., R or Python).

### **A brief tutorial**

In the following section, we go through the necessary steps to use face classification APIs (see Figure A.1 for an overview). We will describe each step together with some tips and tricks, but without programming code. A detailed tutorial on how to use APIs, including code, can be found in the Supplemental Materials and some code examples can be found in Appendix A and B.

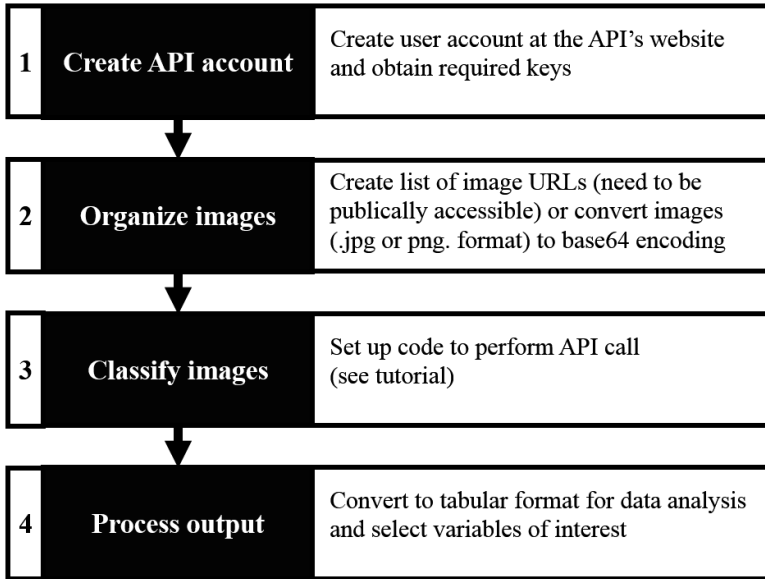


Figure A.1. Overview of the basic steps required for using face classification APIs.

The first step is to obtain the API keys. To obtain the keys, create an account at the website of the API classification service of your choice. It may be necessary to specify a name for an application. Many APIs are aimed at software developers, rather than researchers, which is why they often request information such as the name of the application a developer may be working on. This is simply a label, so any name suffices. After creating the account, and optionally the application, the keys can be obtained. Two keys are necessary: a public key and a secret key. The public key should be readily visible after the account has been created, while the secret key may need to be generated or explicitly revealed. It is important to keep the API keys safe. The API keys represent you as the user and can be used by others to use your account for their purposes, potentially accumulating a substantial amount of processing fees. Take particular heed of this when sharing your code as this is likely to contain your API keys.

The second step is to organize the images to be classified. The images should either be locally stored image files or a list of URLs. Local images (i.e., images stored on your computer) should be in either a .jpg or .png file format and should be converted to a base64 encoding. This can be achieved relatively easily using code (see Appendix A and B for an example using the `base64encode()` function in R). When a list of URLs is used, the URLs must refer to publically accessible images.

The third step is to perform the API call. In order to perform an API call, an HTTP POST request must be made. A POST request consists of sending information to a server and retrieving processed information based on the supplied information. In the case of face classification, we need to supply the public and secret API keys, the image (either a URL or base64 encoded image), and the encoding of the API call itself. The encoding of the API call depends on which face classification service you use. The documentation should include the kind of encoding that is required. Face++ requires a multipart encoding while Kairos requires JSON encoding. Some services also require additional information, such as what kind of face attributes to return. This is the case for the Kairos API, which requires the specification of face attributes such as gender and age to be explicitly specified. For other APIs, such as Face++, this may be optional. The API documentation should indicate what is required and what is optional.

After performing the API call, we recommend to check whether the API call was successfully run. APIs return a status code that can be used to determine whether the call was completed successfully or whether an error occurred. If an error occurred, the status code often provides some information as to what went wrong. For example, it may be the case that the keys are incorrect, that the image file was too large, or that some of the supplied information (often referred to as arguments) was incorrectly specified or missing. In the case of Face++ and Kairos, a successful API call should return the value 200.

The final step is to process the returned data. The Face++ and Kairos API calls return data in JSON format. This data is organized, but not necessarily suitable for data analysis. Preferably, the data is converted to tabular data so that it can easily be merged with other data (e.g., outcome variables or additional face attributes gathered through other means) and used for data analysis. Our detailed tutorial in the Supplemental Material contains an example of R code to convert the result of an API call to tabular data.

It may be fruitful to write code that performs the previous two steps repeatedly. An important consideration in using such a loop is how to handle unsuccessful API calls. Unsuccessful API calls should not break the loop (thus stopping the collection of data) and should also be saved so that it is clear how many images could not be classified. Not all images are suitable for face classification APIs, such as those that have insufficient quality, images with very small faces, or images containing faces that are rotated, thereby hiding a substantial portion of the face, may not result in any classifications. Face classification APIs differ in the extent to which they can effectively process these images of varying quality.

In the following, we present two studies that tested the accuracy of the Face++ algorithm and the Kairos algorithm in classifying a target's gender, age, and race. All data, materials, and scripts are available at Open Science Framework (<https://osf.io/23pn4>).<sup>33</sup>

### **Study A.1**

In Study A.1, we drew on two open-access face databases, the Chicago Face Database (D. S. Ma et al., 2015) and the 10k Faces Database (Bainbridge, Isola, Blank, et al., 2013; Bainbridge, Isola, & Oliva, 2013), to test the algorithms' accuracy.

---

<sup>33</sup> These scripts also contain an example of a loop that performs API calls for a set of images.



## Methods

### Materials.

**Chicago Face Database.** The Chicago Face Database contains images of 597 individuals taken in a controlled lab environment (D. S. Ma et al., 2015). All targets wore a grey shirt and displayed a neutral facial expression. The accompanying data set includes the self-reported gender and race of all targets. The targets' age was determined by showing each image to 20-131 ( $M = 43.74$ ) participants who were asked to provide an age estimate. Age ratings were then averaged across all participants. The Chicago Face Database is particularly suited as it contains targets with widely varying demographic characteristics. Targets indicated belonging to four different racial groups (33.00% Black, 30.65% White, 18.26% Asian, and 18.09% Hispanic). Approximately half of all targets are female (51.42%) and their rated age ranges from 17 to 56 years ( $M = 28.86$ ,  $SD = 6.30$ ). The self-reported gender and race as well as the rated age serve as our benchmarks.

**10k Faces Database.** While the Chicago Face Database contains images of individuals varying in gender, age, and race, the images were taken under controlled conditions in the lab. However, many images people are exposed to in real life—such as profile photos on Facebook, Twitter, or Airbnb—are highly variable. To provide a more conservative test of the API's performance, we used images from the 10k Faces Database (Bainbridge, Isola, Blank, et al., 2013; Bainbridge, Isola, & Oliva, 2013). The full database contains more than 10,000 face images downloaded from the internet, mostly displaying non-famous people. All images were cropped to an oval shape to eliminate background features and resized to the same height. We focus on a subset of 2,222 images for which demographic data is available. A target's gender, age, and race was determined by showing each image to 12 independent MTurk workers

## Appendix

who categorized the faces on the relevant characteristics.<sup>34</sup> We excluded four targets with missing age data and two targets whose race was classified as ‘other’. Our final data set contained 2,216 images. Targets varied in race (82.67% White, 9.93% Black, 4.15% Asian, 3.24% Hispanic) and age (11.10% younger than 20 years, 37.77% 20-30 years old, 31.68% 30-45 years old, 17.64% 45-60 years old, 1.81% older than 65 years). There were slightly more men than women (42.69% female). The ratings provided by MTurk workers served as our benchmark.

**Procedure and analysis plan.** We used the Face++ API and the Kairos API to classify the gender, age, and race of all targets. Kairos provides confidence estimates for each gender and race category and we selected the category with the highest confidence estimate as Kairos’ classification output. For each dimension, we compared the API’s classification against the database-specific benchmark to determine the algorithm’s accuracy. For the Chicago Face Database, the benchmark is the target’s self-reported gender and race, as well as the average age estimate provided by human raters. For the 10k Faces Database, the benchmark is the gender, age, and race of targets as classified by human raters.

To estimate the performance of the APIs, we calculated their sensitivity, specificity, and accuracy (Baratloo, Hosseini, Negida, & El Ashal, 2015). These estimates are based on the number of true positives (TP; e.g., a White individual classified as White), false positives (FP; e.g., a non-White individual classified as White), true negatives (TN; e.g., a non-White individual classified as non-White), and false negatives (FN; e.g., a White individual classified as non-White). *Sensitivity* denotes the percentage of actual occurrences that were detected by the algorithm. For example, a sensitivity of 90% for classifying White targets means

---

<sup>34</sup> A target’s age was determined by taking the average estimated age across the 12 raters. Targets were then categorized into one of five age groups. A target’s gender and race were determined by taking the modal response of raters.

that 90% of all White targets were also classified as such. Sensitivity is calculated by dividing the number of true positives by the total number of targets:  $\frac{TP}{TP+FN}$ . *Specificity* denotes the percentage of detected occurrences that reflect actual occurrences. For example, a specificity of 90% for classifying White targets means that of all targets that were classified as White, 90% are actually White. Specificity is calculated by dividing the number of true positives by the total number of targets:  $\frac{TN}{TN+FP}$ . *Accuracy* represents the algorithm's overall ability to discriminate between targets (e.g., accurately classifying their race) and is calculated by dividing the sum of true positives and true negatives by the total number of targets:  $\frac{TP+TN}{TP+TN+FP+FN}$ .

## Results

Before analyzing the classification accuracy of the algorithms, we tested if the algorithms were able to detect a face and thus provide a classification for every image. Both Face++ and Kairos detected a face in all 597 images of the Chicago Face Database. For the more variable images of the 10k Faces Database, Face++ detected a face in all 2,216 images while Kairos detected a face in 2,208 images (99.64%). Thus, the face detection rate of both algorithms was close to 100%. The results reported here are based on all images for which both algorithm were able to provide a classification.

**Gender.** We first compared the gender the Face++ algorithm assigned to a given target with the benchmark gender of the targets from both databases (see Table A.1). Accuracy was at 88.94% for the Chicago Face Database, 95% confidence interval (CI) [86.15%, 91.35%] and at 90.17% for the 10k Faces Database, 95% CI [88.85%, 91.38]. Accuracy levels did not significantly differ between the two samples,  $\chi^2(1) = 0.65$ ,  $p = .42$ ,  $\Delta = 1.23\%$ . Thus, we did not find any evidence that the

## Appendix

performance of the Face++ algorithm in classifying gender was lower for the more variable image set.

Next, we compared the gender the Kairos algorithm assigned to a given target with the benchmark gender of the targets from both databases (see Table A.1). Accuracy was at 96.15% for the Chicago Face Database, 95% CI [94.28%, 97.54%] and at 98.55% for the 10k Faces Database, 95% CI [97.96%, 99.01]. Surprisingly, performance was slightly better for the more variable image set,  $\chi^2(1) = 12.86$ ,  $p < .001$ ,  $\Delta = 2.40\%$ .

Finally, we compared the performance of the two algorithms. The Kairos algorithm was more accurate than the Face++ algorithm when classifying faces from both the Chicago Face Database (7.21 percentage points difference,  $\chi^2(1) = 22.46$ ,  $p < .001$ ) and from the 10k Faces Database (8.38% difference,  $\chi^2(1) = 144.11$ ,  $p < .001$ ). In sum, the Kairos algorithm showed better performance in gender classification for both controlled and more variable face images.

**Table A.1**  
Accuracy of the Face++ algorithm in classifying the gender of targets from the Chicago Face Database and the 10k Faces Database (Study A.1).

	Chicago Face Database		10k Faces Database	
	Female	Male	Female	Male
<b>Sensitivity</b>				
Face++	82.08%	96.21%	88.41%	91.50%
Kairos	93.16%	99.31%	98.10%	98.89%
<b>Specificity</b>				
Face++	96.21%	82.08%	91.50%	88.41%
Kairos	99.31%	93.16%	98.89%	98.10%
<b>Accuracy</b>				
Face++	88.94% [86.15%, 91.35%]		90.17% [88.85%, 91.38%]	
Kairos	96.15% [94.28%, 97.54%]		98.55% [97.96%, 99.01%]	

## Appendix

**Age.** To test the algorithms' accuracy in age classification, we first compared the age the Face++ algorithm assigned to a given target with the benchmark age of the targets from both databases (i.e., the error in age estimation). For the Chicago Face Database targets, the average error for estimated age was 7.98 years ( $SD = 5.67$ ), which is significantly different from zero,  $t(596) = 34.38, p < .001$  (Figure A.2A). We also compared the age Face++ assigned to a given target with the benchmark age of the 10k Faces Database targets. Figure A.2A shows that that the average age estimated by Face++ shifted upwards with each age category. We calculated the percentage of age estimates that fell within the benchmark age category. Across the five age categories, only 18.34% of age estimates fell within the benchmark age range. Examining the distance between targets' assigned age category and their benchmark age category showed that for the majority of targets, age estimates were only off by one category ( $M = 1.11, SD = 0.73$ ).

Next, we compared the age the Kairos algorithm assigned to a given target with the benchmark age of the targets from both databases. For the Chicago Face Database targets, the average error for estimated age was 3.30 years ( $SD = 2.64$ ), which is significantly different from zero,  $t(596) = 30.58, p < .001$  (Figure A.2B). We also compared the age Kairos assigned to a given target with the benchmark age of the 10k Faces Database targets. Figure A.2B shows that that the average age estimated by Kairos shifted upwards with each age category. We calculated the percentage of age estimates that fell within the benchmark age category. Across the five age categories, 38.95% of age estimates fell within the benchmark age category. Examining the distance between targets' benchmark age category and their assigned age category showed that for the majority of targets, age estimates were only off by one category ( $M = 0.65, SD = 0.56$ ).

Finally, we compared the performance of the two algorithms. For the Chicago Face Database targets, the Kairos algorithm was significantly

more accurate than the Face++ algorithm, with an average difference in error for estimated age of 4.68 years,  $t(842.46) = 18.28$ ,  $p < .001$ . For the 10k Faces Database, the majority of age estimates of both algorithms fell outside of the benchmark age category (Face++: 81.66%, Kairos: 61.05%). However, age estimates of the Kairos algorithm were significantly more often within this age range,  $\chi^2(1) = 228.42$ ,  $p < .001$ ,  $\Delta = 20.61\%$ . Moreover, the mean distance between a target's estimated age category and their benchmark age category was smaller for the Kairos algorithm,  $t(4124.1) = 23.59$ ,  $p < .001$ ,  $\Delta = 0.46$ . In sum, our findings show that age estimates by Kairos were more accurate for the controlled and more variable face images.

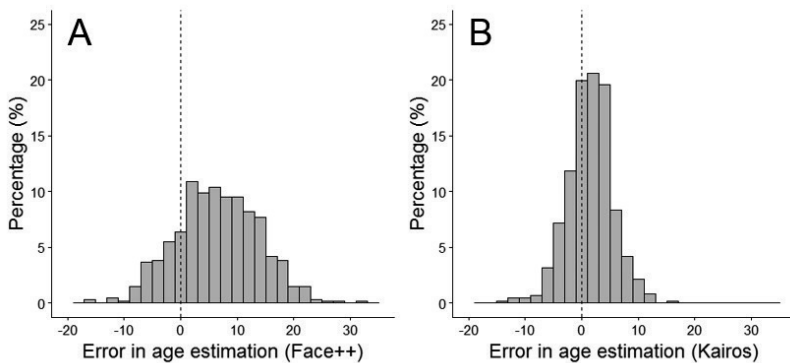
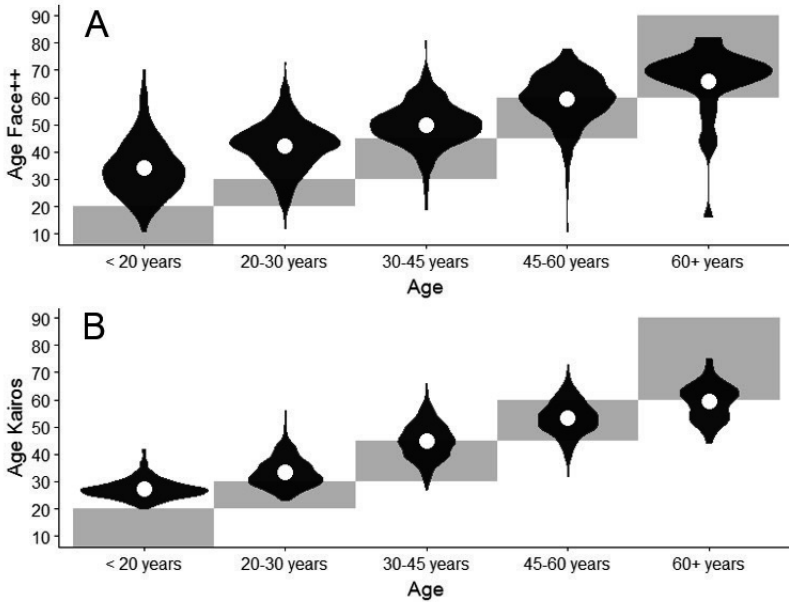


Figure A.2. Distribution of the difference between the age estimated by (A) the Face++ algorithm or (B) the Kairos algorithm and the average age estimate of human raters for the Chicago Face Database (Study A.1). The dashed line represents no difference between the algorithm and human raters. Observations left of the dashed line represent an underestimation by the algorithm whereas observations right of the dashed line represent an overestimation by the algorithm.

## Appendix



*Figure A.3.* The distribution of estimated age by (A) Face++ or (B) Kairos as a function of benchmark age category (Study A.1). White dots denote the average estimated age of the algorithm. The shaded areas illustrates the targets' benchmark age category. The overlap between the age distribution and the shaded area represents the proportion of age estimates by the algorithm that fell within the targets' benchmark age category.

**Race.** To test the algorithms' accuracy in race classification, we first compared the race the Face++ algorithm assigned to a given target with the benchmark race of the targets from both databases. Accuracy was at 72.86%, 95% CI [69.11%, 76.39%] for the Chicago Face Database and at 82.79%, 95% CI [81.15%, 84.34%] for the 10k Faces Database (Table A.2). There was a significant difference in accuracy levels between the two samples,  $\chi^2(1) = 29.09$ ,  $p < .001$ ,  $\Delta = 9.93\%$ , showing that the



performance of the Face++ algorithm in classifying race was better for the 10k Faces Database.<sup>35</sup>

Next, we compared the race the Kairos algorithm assigned to a given target with the benchmark race of the targets from both databases. Accuracy was at 89.28%, 95% CI [86.52%, 91.65%] for the Chicago Face Database and at 95.06%, 95% CI [94.08, 95.93] for the 10k Faces Database (Table A.3). Accuracy levels differed significantly between the two samples,  $\chi^2(1) = 26.13$ ,  $p < .001$ ,  $\Delta = 5.78\%$ , showing that the algorithm's performance was better for the 10k Faces Database.

Finally, we compared the performance of the two algorithms for both databases. Results showed that the Kairos algorithm outperformed the Face++ algorithm by 16.42 percentage points for the Chicago Face Database,  $\chi^2(1) = 51.38$ ,  $p < .001$ , and by 12.27 percentage points for the 10k Faces Database,  $\chi^2(1) = 167.52$ ,  $p < .001$ .

---

<sup>35</sup> Face++ does not provide a classification for Hispanics. Accuracy was at 88.96%, 95% CI [85.84%, 91.59%] for the Chicago Face Database and at 85.58%, 95% CI [84.02, 87.04] for the 10k Faces Database when we focused only on non-Hispanic targets.

## Appendix

Table A.2

Accuracy of the Face++ algorithm in classifying the race of targets from the Chicago Face Database and the 10k Faces Database (Study A.1).

	Asian	Black	Hispanic	White
<b>Sensitivity</b>				
Chicago	90.83%	90.86%	-	85.79%
10k	64.13%	75.91%	-	87.83%
<b>Specificity</b>				
Chicago	86.27%	92.25%	-	84.54%
10k	92.68%	94.12%	-	71.88%
<b>Accuracy</b>				
Chicago	72.86% [69.11%, 76.39%]			
10k	82.79% [81.15%, 84.34%]			

Table A.3

Accuracy of the Kairos algorithm in classifying the race of targets from the Chicago Face Database and the 10k Faces Database (Study A.1).

	Asian	Black	Hispanic	Other	White
<b>Sensitivity</b>					
Chicago	93.58%	94.42%	66.67%	-	94.54%
10k	73.91%	95.00%	59.72%	-	97.53%
<b>Specificity</b>					
Chicago	98.36%	98.75%	96.29%	-	93.38%
10k	99.34%	99.65%	97.47%	-	93.49%
<b>Accuracy</b>					
Chicago	89.28% [86.52%, 91.65%]				
10k	95.06% [94.08%, 95.93%]				

## Discussion

Results of Study A.1 showed relatively high levels of classification accuracy for all three demographic characteristics. Kairos correctly classified the gender of approximately 98% of targets and the race of 94% of targets. Face++'s performance was slightly lower, with 90% correct gender classifications and 80% correct race classifications. Lower performance on race classification was partly due to the fact that Face++ does not detect Hispanic targets and all Hispanic targets in our data sets were consequently misclassified. Accuracy improved to 86% when restricting our analyses to non-Hispanic targets, but was still below the accuracy level of Kairos. In general, classification accuracy of both algorithms varied depending on the race of the target. For example, Kairos correctly classified 98% of all White targets from the 10k Faces database, but only 60% of all Hispanic targets. Face++ correctly classified 88% of all White targets, but only 64% of all Asian targets. Finally, both algorithms tended to overestimate the age of targets. However, across both databases, the average error in estimated age was lower for Kairos.

In sum, both algorithms provided a relatively accurate classification of demographic characteristics based on controlled and variable face images. Accuracy often depended on the specific category being classified. Neither algorithm showed decreased performance on any characteristic when face images were not taken in highly standardized conditions but were more variable regarding image quality, lighting condition, head pose, and facial expression.

## Study A.2

In Study A.2, we investigated whether the algorithms' performance is affected by two factors: the facial expression of targets and the angle from which targets were photographed. We also recruited human participants to classify images. Our goal was to test if any effect of facial expression or camera angle on the algorithms' classification accuracy

## Appendix

similarly influences the accuracy of human raters. To this end, we assessed the classification accuracy of human raters, the Face++ algorithm, and the Kairos algorithm as a function of camera angle (frontal vs. three-quarter profile) and facial expressions of targets (neutral vs. smiling). Thus, the present study was designed to show (a) under which conditions the algorithms' performance might decrease and (b) whether the same is true for human raters.

### Methods

**Participants.** We recruited 186 participants ( $M_{\text{age}} = 35.03$ ,  $SD_{\text{age}} = 10.18$ ; 38.17% female) from Amazon Mechanical Turk (Paolacci & Chandler, 2014) to classify demographic characteristics in exchange for \$0.80 each. Each participant was assigned to one of four image sets (frontal-neutral, frontal-smiling, profile-neutral, or profile-smiling) and to one of three rating tasks (age, gender, or race). We ensured that each image was assessed by at least 15 raters.

**Materials.** We used the Facelab London Database which contains images of 102 individuals and self-report data for targets' gender, age, and race (DeBruine & Jones, 2017). For the present study, we excluded one biracial individual and two individuals who did not report their age. The remaining 99 targets indicated belonging to three different racial groups (68.69% White, 18.18% Asian, and 13.13% Black). Approximately half of targets are female (48.48%) and their ages range from 18 to 54 ( $M = 27.79$ ,  $SD = 7.11$ ). The self-reported demographics served as our benchmark. The database contains several different image sets. Here, we focused on four different sets for which the facial expression of the target (neutral vs. smiling) and the angle from which the photo was taken (frontal vs. three-quarter profile) were varied.

**Procedure.** Participants were shown one image at a time in a random order. Depending on the condition they were asked to guess the person's gender (by selecting male or female), age (by dragging a slider

that ranged from 1 to 100), or race (by selecting White, Black, Asian, or Hispanic). Participants only saw images from one of the four image sets. The rated gender and race of a target were determined by taking participants' modal response (i.e., the option that was chosen most often). The rated age of a target was determined by averaging the estimates of all raters. Finally, we again used the Face++ algorithm and the Kairos algorithm to classify the targets' gender, age, and race for each image set.

## Results

Before analyzing classification accuracy, we again tested whether the algorithms were able to detect a face and thus provide a classification for every image. Face++ detected a face in all 396 images (four stimulus sets of 99 targets). Kairos detected a face in all frontal-neutral and frontal-smiling images. It did not detect 14 targets (14.14%) in the profile-neutral image set and five targets (5.05%) in the profile-smiling image set. Overall, 19.19% of targets were not detected by Kairos. Because our goal was to directly compare the accuracy of the algorithms, the results reported here are based on 84 targets for which both algorithms were able to provide a classification across all four image sets ( $n = 336$ ).

**Gender.** To compare accuracy in gender classification, we first tested how well each classification method performed across the four image sets. Accuracy was at 100.00% for human raters, at 91.07%, 95% CI [87.50%, 93.89%] for Face++, and at 99.11%, 95% CI [97.41%, 99.82%] for Kairos. Regressing accuracy on classification method showed that accuracy of the Face++ algorithm was significantly lower than accuracy of human raters,  $b = -4.204$ ,  $SE = 1.430$ ,  $\chi(2) = 39.26$ ,  $p < .001$ . Accuracy of the Kairos algorithm was not significantly lower than accuracy of human raters,  $b = -1.955$ ,  $SE = 1.516$ ,  $\chi(2) = 2.54$ ,  $p = .11$ . We also found that accuracy of the Kairos algorithm was significantly higher

## Appendix

than accuracy of the Face++ algorithm,  $b = 2.249$ ,  $SE = 0.571$ ,  $\chi(2) = 25.73$ ,  $p < .001$ . Thus, we found that Face++, but not Kairos, was outperformed by human raters in gender classification. Performance of the Kairos algorithm was significantly better than performance of the Face++ algorithm.

Next, we tested how the classification accuracy of human raters and of the two algorithms was affected by camera angle (coded -0.5 for frontal and +0.5 for three-quarter profile) and facial expression (coded -0.5 for neutral and +0.5 for smiling; see Figure A.4). Human raters showed perfect levels of accuracy across the four image sets. Thus, we did not find any evidence that their performance was affected by camera angle or facial expression.

For Face++, we found a negative effect of camera angle,  $b = -1.604$ ,  $SE = 0.476$ ,  $\chi(3) = 14.67$ ,  $p < .001$ , showing that accuracy was lower for targets photographed from three-quarter profile (85.12%) as compared to targets photographed from the front (97.02%). We found no effect of facial expression,  $b = -0.038$ ,  $SE = 0.476$ ,  $\chi(3) = 0.01$ ,  $p = .94$ , and no interaction effect between camera angle and facial expression,  $b = 0.622$ ,  $SE = 0.951$ ,  $\chi(3) = 0.44$ ,  $p = .51$ . Thus, Face++'s performance was affected by camera angle, but not by targets' facial expression.

For Kairos we found no effect of camera angle,  $b = -0.991$ ,  $SE = 1.266$ ,  $\chi(3) = 0.62$ ,  $p = .43$ , no effect of facial expression,  $b = -0.991$ ,  $SE = 1.266$ ,  $\chi(3) = 0.62$ ,  $p = .43$ , and no interaction effect between camera angle and facial expression,  $b = -1.982$ ,  $SE = 1.532$ ,  $\chi(3) = 0.62$ ,  $p = .43$ . Thus, we did not find any evidence that Kairos' performance was affected by camera angle or by facial expression.

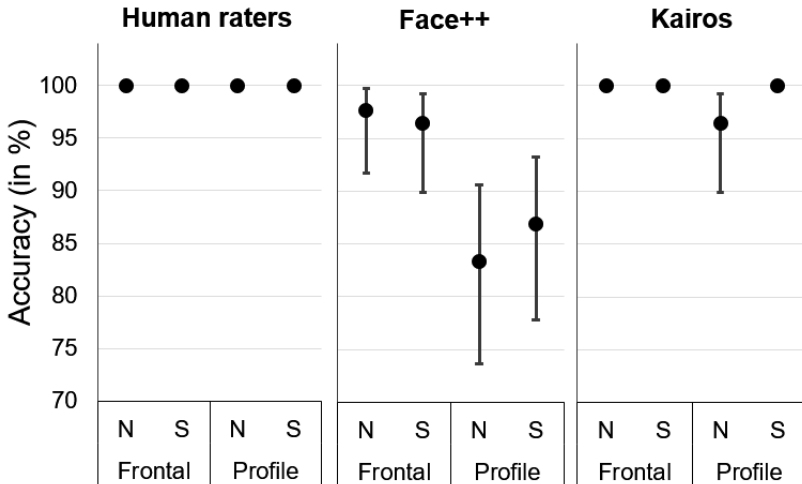


Figure A.4. Accuracy of gender classifications provided by human raters, the Face++ algorithm, and the Kairos algorithm as a function of facial expression (“N” = neutral, “S” = smiling) and camera angle (Study A.2).

**Age.** To compare accuracy in age classification, we first tested how well each classification method performed across the four image sets. We again computed the average difference between estimated age and actual age as an indicator of the error in age classification. Results showed an error of 4.59 years ( $SD = 2.77$ ) for human raters, 13.93 years ( $SD = 7.76$ ) for Face++, and 4.24 years ( $SD = 2.99$ ) for Kairos. Regressing error in estimated age on classification method showed that the error of the Face++ algorithm was significantly larger than the error of human raters,  $b = 9.333$ ,  $SE = 0.390$ ,  $t(1005) = 23.91$ ,  $p < .001$ . On the other hand, the error of the Kairos algorithm was not significantly larger than the error of human raters,  $b = -0.348$ ,  $SE = 0.390$ ,  $p = .37$ . In turn, the error of Face++ was larger than the error of Kairos,  $b = 9.682$ ,  $SE = 0.390$ ,  $p < .001$ . Thus, we found that Face++, but not Kairos, was outperformed by human raters in age classification. Performance of the Kairos algorithm was significantly better than performance of the Face++ algorithm.

## Appendix

Next, we tested how error in age classification of human raters and of the two algorithms was affected by camera angle and facial expression (see Figure A.5). For human raters, we found no effect of camera angle,  $b = -0.451$ ,  $SE = 0.289$ ,  $t(332) = 1.56$ ,  $p = .12$ , a negative effect of facial expression,  $b = -0.919$ ,  $SE = 0.289$ ,  $t(332) = 3.18$ ,  $p = .002$ , and a significant interaction effect between camera angle and facial expression,  $b = -2.747$ ,  $SE = 0.578$ ,  $t(332) = 4.75$ ,  $p < .001$ . Results showed that error in estimated age was smaller for smiling targets ( $M = 3.22$ ,  $SD = 2.22$ ) as compared to neutral targets ( $M = 5.51$ ,  $SD = 2.90$ ) when they were photographed from three-quarter profile,  $b = -2.292$ ,  $SE = 0.409$ ,  $t(322) = 5.61$ ,  $p < .001$ . For targets photographed from the front, there was no significant difference between smiling ( $M = 5.05$ ,  $SD = 2.79$ ) and neutral ( $M = 4.59$ ,  $SD = 2.64$ ) targets,  $b = 0.455$ ,  $SE = 0.409$ ,  $t(322) = 1.11$ ,  $p = .27$ . Thus, performance of human raters was affected by the facial expression of targets, but only when targets were photographed from three-quarter profile. Specifically, error in age classification was smaller for smiling targets.

For Face++, we found a positive effect of camera angle,  $b = 4.756$ ,  $SE = 0.800$ ,  $t(332) = 5.94$ ,  $p < .001$ , showing that error in estimated age was larger for targets photographed from three-quarter profile ( $M = 16.30$ ,  $SD = 8.21$ ) as compared to targets photographed from the front ( $M = 11.55$ ,  $SD = 6.48$ ). We also found a positive effect of facial expression,  $b = 2.149$ ,  $SE = 0.800$ ,  $t(332) = 2.69$ ,  $p = .008$ , showing that the error in estimated age was larger for smiling targets ( $M = 15.00$ ,  $SD = 7.96$ ) than for neutral targets ( $M = 12.85$ ,  $SD = 7.42$ ). There was no significant interaction effect between camera angle and facial expression,  $b = -0.893$ ,  $SE = 1.601$ ,  $t(332) = 0.56$ ,  $p = .58$ . Thus, performance of the Face++ algorithm was affected by camera angle and facial expression with larger error in estimated age for targets photographed from three-quarter profile and for smiling targets.



For Kairos, we found no effect of camera angle,  $b = 0.048$ ,  $SE = 0.327$ ,  $t(332) = 0.15$ ,  $p = .88$ , no effect of facial expression,  $b = 0.310$ ,  $SE = 0.327$ ,  $t(332) = 0.95$ ,  $p = .35$ , and no interaction effect between camera angle and facial expression,  $b = -0.167$ ,  $SE = 0.654$ ,  $t(332) = 0.26$ ,  $p = .80$ . Thus, we did not find any evidence that performance of the Kairos algorithm was affected by camera angle or by facial expression.

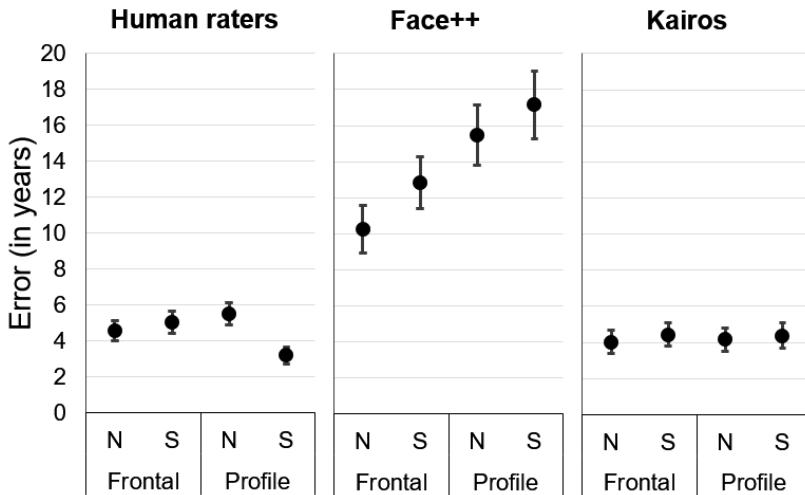


Figure A.5. Accuracy of age classifications provided by human raters, the Face++ algorithm, and the Kairos algorithm as a function of facial expression (“N” = neutral, “S” = smiling) and camera angle (Study A.2).

**Race.** To compare accuracy in race classification, we first tested how well each classification method performed across the four image sets. Accuracy was at 88.99%, 95% CI [85.14%, 92.13%] for human raters, at 72.02%, 95% CI [66.90%, 76.76%] for Face++, and at 77.68%, 95% CI [72.84%, 82.02%] for Kairos. Regressing accuracy on classification method showed that accuracy of the Face++ algorithm was significantly lower than accuracy of human raters,  $b = -1.135$ ,  $SE = 0.212$ ,  $\chi(2) = 31.45$ ,  $p < .001$ . Accuracy of the Kairos algorithm was also significantly lower than accuracy of human raters,  $b = -0.835$ ,  $SE = 0.212$ ,

## Appendix

$\chi(2) = 15.61, p < .001$ . Accuracy of the Kairos algorithm was slightly higher than accuracy of the Face++ algorithm but this difference was only marginally significant,  $b = -0.300, SE = 0.179, \chi(2) = 2.84, p = .092$ . Thus, we found that Face++ and Kairos were outperformed by human raters in race classification. We found no significant difference in performance between Face++ and Kairos.

Next, we tested how the classification accuracy of human raters and of the two algorithms was affected by camera angle and facial expression (see Figure A.6). For human raters, we found no effect of camera angle,  $b = 0.171, SE = 0.344, \chi(3) = 0.25, p = .62$ , no effect of facial expression,  $b = -0.047, SE = 0.344, \chi(3) = 0.02, p = .89$ , and no interaction effect between camera angle and facial expression,  $b = -0.342, SE = 0.688, \chi(3) = 0.25, p = .62$ . Thus, we did not find any evidence that performance of human raters was affected by camera angle or facial expression.

For Face++, we found a negative effect of camera angle,  $b = -1.498, SE = 0.271, \chi(3) = 34.60, p < .001$ , showing that accuracy was lower for targets photographed from three-quarter profile (57.74%) as compared to targets photographed from the front (86.31%). We found no effect of facial expression,  $b = 0.121, SE = 0.271, \chi(3) = 0.20, p = .65$ , and no interaction effect between camera angle and facial expression,  $b = 0.048, SE = 0.543, \chi(3) = 0.01, p = .93$ . Thus, performance of the Face++ algorithm was affected by camera angle, but not by facial expression, with lower performance for targets photographed from three-quarter profile.

For Kairos, we found a negative effect of camera angle,  $b = -1.320, SE = 0.291, \chi(3) = 23.06, p < .001$ , showing that accuracy was lower for targets photographed from three-quarter profile (66.67%) as compared to targets photographed from the front (88.69%). We found no effect of facial expression,  $b = 0.209, SE = 0.291, \chi(3) = 0.52, p = .47$ , and no interaction effect between camera angle and facial expression,  $b = 0.645, SE = 0.582, \chi(3) = 1.24, p = .26$ . Thus, performance of the Kairos algorithm

was affected by camera angle, but not by facial expression, with lower performance for targets photographed from three-quarter profile.

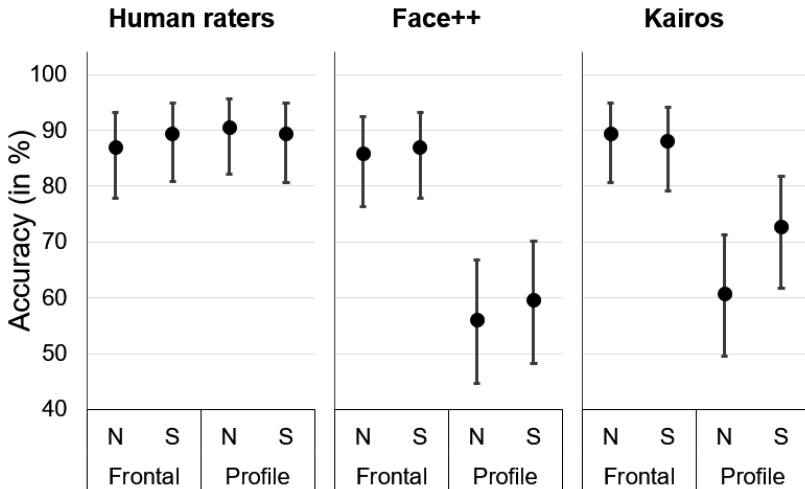


Figure A.6. Accuracy of race classifications provided by human raters, the Face++ algorithm, and the Kairos algorithm as a function of facial expression (“N” = neutral, “S” = smiling) and camera angle (Study A.2).

## Discussion

Results of Study A.2 extended our previous findings in three important ways. First, we estimated the accuracy of the two algorithms for a new set of images and compared their performance in gender, age, and race classification directly to the performance of human raters. Human raters consistently outperformed the Face++ algorithm for all three characteristics. However, the Kairos algorithm was only outperformed by human raters in the classification of race.

Second, assessing classification accuracy as a function of camera angle from which targets were photographed (frontal vs. three-quarter profile) and their facial expression (neutral vs. smiling) showed which image properties might influence the performance of the two algorithms and human raters. Neither factor had an impact on the accuracy of

human raters. However, accuracy in gender, age, and race classification of the Face++ algorithm was reduced for targets photographed from three-quarter profile. Smiling only negatively influenced accuracy in age classification. For the Kairos algorithm, accuracy in race, but not gender or age classification was reduced for images photographed from three-quarter profile. Facial expression did not influence accuracy for any characteristic. In sum, our results revealed two conditions under which the performance of algorithms, but not that of human raters might suffer. However, the exact decrease in accuracy is dependent on which characteristic is classified and which algorithm is used.

Third, one shortcoming of our first study was that we were only able to compare age estimates of the two algorithms against age estimates of human raters rather than self-reported age. The current study compared the algorithms' age estimates against self-reported age and confirmed that error is larger for Face++ than for Kairos.

### **General discussion**

Many important social outcomes are shaped by a person's gender, age, or race and exploring the influence of demographic characteristics has been a topic of intense study across the social sciences and in psychology in particular. With more social interactions moving to online environments where profile photos are prevalent (e.g., economic exchange, dating, and social networking), new methods for data extraction (Landers, Brusso, Cavanaugh, & Collmus, 2016), and a general increase in the availability of data relevant for social scientists (E. E. Chen & Wojcik, 2016; Kosinski, Wang, Lakkaraju, & Leskovec, 2016; Lazer et al., 2009), researchers are afforded the opportunity to study the influence of demographic characteristics using large naturalistic data sets. Given these developments, automated face classification can be a useful tool. We presented a tutorial and R code on how to use two face classification algorithms and tested their performance by drawing on

three face databases ( $n = 3,141$  images). Our results show that the algorithms' accuracy is generally high and close to the accuracy level of human raters.

### **Evaluating and comparing the algorithms' performance**

While we have argued that face classification algorithms can be an effective alternative to human raters, their usefulness ultimately depends on their accuracy. Across the various tests presented here, we found that the algorithms' accuracy depended on a variety of factors such as the targets' race, the angle from which targets were photographed, and which algorithm was used. However, three general conclusions can be drawn from our results. First, our tests showed that Kairos outperformed Face++ in classifying all three demographic characteristics.

Second, the Kairos algorithm's accuracy levels are generally high and close to those of human raters. The Kairos algorithm performed as well as human raters, with the exception being race classification. Yet, lower accuracy in race classification was only found for targets photographed from three-quarter profile (as opposed to the front). Thus, we conclude that algorithms can provide accurate classifications along the demographic dimensions of gender, age, and race.

Third, the accuracy of the algorithms depended on various characteristics of the images. Our studies were designed to test classification accuracy under varying conditions. Results of Study A.1 showed no indication that accuracy was reduced for more variable (10k Faces Database) rather than standardized images (Chicago Face Database). In other words, accuracy levels were similar even when properties such as camera angle, facial expression, head tilt, image quality, and lighting conditions were not controlled. This observation is important as many potential data sets of interest contain variable photos, such as profile photos on Airbnb (Edelman et al., 2017) or

screenshots of TV game show footage (Darai & Grätz, 2013). Study A.2 showed that the accuracy levels were affected by the angle from which a targets were photographed. Compared to human raters, the Kairos algorithm was less accurate in classifying the race of targets photographed from three-quarter profile, as opposed to the front. However, camera angle did not influence accuracy in gender or age classification. Accuracy was also unaffected by targets' facial expression (neutral vs. smiling).

Thus, we found some evidence that unlike human raters, the Kairos algorithm's accuracy in race, but not gender or age classification was lower for targets photographed from three-quarter profile (though performance was unaffected by the facial expression of targets). Taken together, our findings demonstrate that algorithms can provide accurate classifications of demographic characteristics, even for variable, non-standardized images downloaded from the internet. In fact, we find little evidence that the performance of human raters was better than that of the algorithms (especially the Kairos algorithm).

### **Advantages and limitations of using face classification APIs**

Relying on automated face classification procedures rather than human participants has several key advantages. Next to obvious benefits for individual researchers, such as less time spent on data collection, we want to highlight two more general advantages. With automated classification, a researcher's sample size is no longer limited by the size of their participant pool and, to a much lesser extent, by their research budget. This means that hypotheses can be tested using large sample sizes, providing high statistical power. By definition, studies with high statistical power will detect true relationships more often, thus reducing the number of false negatives in the literature. Research lines with high statistical power also produce more accurate effect size estimates and a higher proportion of statistically significant results that actually reflect

true relationships (Button et al., 2013; Ioannidis, 2005). In sum, high statistical power is essential for producing reliable research and recent large-scale failures to replicate established findings in psychology have led to an increased focus on power (Fraleay & Vazire, 2014; Open Science Collaboration, 2015).

We also hope that the availability of easily accessible APIs will encourage researchers to test their hypotheses using large, naturalistic data sets. While studies from both the lab and the field are needed to convincingly demonstrate an effect, scholars have noted that the latter is often neglected by psychologists, calling for more studies that analyze real-world data (Baumeister et al., 2007; Maner, 2016). This call coincides with an increasing availability of large data sets that can be used to test psychological theories (E. E. Chen & Wojcik, 2016; Kosinski et al., 2016). For scholars interested in studying the effects of age, gender, and race, automated face classification makes large, naturalistic data sets more accessible for research.

Relying on commercial software also has potential drawbacks. It is often unclear how algorithms operate and what data sets they were trained on. Therefore, it is crucial to rigorously test and validate algorithms before they are used in research. We provided first evidence for their validity here, but future studies need to test the algorithms under different conditions. For example, while we tested the algorithms' accuracy in classifying variable images taken from the internet, future studies should look at accuracy levels for profile photos from Facebook or Airbnb, which have been used in recent research (Edelman et al., 2017; Kosinski, 2017). Future studies should also test the algorithms' performance for a wider range of race categories as well as for biracial individuals.

## **Ethical considerations**

Conducting studies with naturalistic data sets—a context in which face classification algorithms are particularly useful—presents unique challenges to researchers who have to ensure that ethical standards are met. At the current moment, there is no comprehensive set of guidelines determining when and how online data can be ethically used, and standards may vary between different institutional review boards (IRB; Chen & Wojcik, 2016; Michal Kosinski, Matz, Gosling, Popov, & Stillwell, 2015). However, this should not be taken as an excuse to dismiss ethical considerations altogether. Researchers should consult their IRB to ensure that their study conforms to local ethical guidelines. Three points in particular deserve special consideration.

First, many studies in computational social science rely on data that is public, but was not created for research purposes (e.g., ebay listings, social network activity). This can make it difficult or even impossible to obtain informed consent from individuals providing the data. Some have argued that public data on the internet should be treated as archival data, which can be used without acquiring informed consent (Kosinski et al., 2015). Given the lack of clear guidelines, researchers can ask themselves how likely it is that people would object to the use of their data. While a researcher's evaluation might not be objective or unbiased, there are differences in the sensitivity of data sets that most people probably agree on. For example, a person's number of followers on a social networking site or the price of an item they are selling in a peer-to-peer market is easily accessible to a large audience and widely disseminating this information is often the central aim of the website's user. Other types of data are more sensitive. When someone discloses their sexual preferences on a dating website or discusses controversial topics in a chat room, this information is only addressed to a very specific audience. People might be more likely to object to information being recorded by a researcher if it relates to sensitive issues and was never meant to be



widely publicized. If a study deals with such data, attempts could be made to obtain informed consent from the relevant individuals.

A related issue concerns the anonymization of data. Researchers need to ensure that any identifying information is removed when data sets are collected or shared. In some contexts, this might be more difficult than anticipated. While it is relatively easy to remove obvious identifiers such as names, addresses, or IP addresses, a person's identity can often be inferred from other information. For example, in the context of Airbnb, it might be possible to identify a host from a combination of data points such as the neighborhood they live in, the size of their apartment, and the price they are asking. Guaranteeing a person's anonymity is a particularly important issue when dealing with personal photos. Just like other personal identifiers, photos should not be shared without the person's consent. Here, relying on an algorithm to classify images can actually help in ensuring anonymity as the images do not have to be shown to human participants in order to collect data on demographic characteristics.

Finally, researchers should explicitly weigh the costs and benefits of conducting a study. Any research design should attempt to minimize harm to participants. As the risk factors of a study increase (e.g., collection of sensitive data, potentially imperfect anonymization of the data), researchers need to critically evaluate the scientific value of their study to assess whether the risks can be justified. In sum, researchers should be aware that ensuring ethical standards in research is particularly challenging but important when dealing with large sets of naturalistic data that individuals did not provide for research purposes. Even if there are no clear restrictions regarding the use of a specific data set, researchers should consult their local IRB to ensure that broader ethical guidelines are met.

### **Practical recommendations**

There are several ways in which the use of face classification algorithms can be optimized. For gender and race classification, Kairos provides confidence estimates for each category. Here, we selected the category with the highest confidence estimate as the detected category. However, researchers can also exclude images that could not be classified with a pre-determined level of confidence. For example, when studying the effect of race in a large data set, researchers could restrict their analyses to images for which the algorithm was able to determine the target's race with at least 90% confidence. Excluding images will lower the sample size, but this might be a price worth paying to reduce error in classifications, especially when the initial data set is very large. Given a large enough sample, we recommend that the robustness of any effect is investigated by varying the confidence threshold for classifications.

At the same time, researchers need to be aware that systematic exclusion of images might introduce selection bias. For example, a researcher interested in racial disparities in living situations might examine whether the apartments of White vs. non-White hosts on Airbnb are located in less desirable neighborhoods. Setting a high confidence threshold for race classifications might lead to more accurate classifications, but also to the exclusion of a considerable number of hosts. This exclusion might not be random. The algorithm's confidence may be lower for Hispanic targets compared to Asian targets. If Hispanic targets live, on average, in less desirable neighborhood than Asian targets, then their exclusion will lead to an underestimation of the true difference between White and non-White hosts.

Finally, researchers should be aware of the characteristics of their image set. As our results have shown, classification accuracy is dependent on several factors. For example, accuracy will be lower—in other words, measurement error will be higher—for Hispanic targets

and for targets photographed from three-quarter profile. Researchers need to manually examine at least a part of their image set to check whether image properties allow for accurate classifications.

## **Conclusion**

Large naturalistic data sets afford researchers to test their theories with high statistical power using data that reflects real-world behavior. For researchers studying the influence of demographic characteristics, this can be a challenge since a large number of participants is needed to classify target's gender, age, or race. The results presented here suggest that face classifications algorithms are often as accurate as human raters. Algorithms are easy-to-use and more time-efficient, therefore providing a useful alternative to human raters.



# References

- Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology, 93*(5), 751–763. <https://doi.org/10.1037/0022-3514.93.5.751>
- Adams, R. B., Nelson, A. J., Soto, J. A., Hess, U., & Kleck, R. E. (2012). Emotion in the neutral face: A mechanism for impression formation? *Cognition & Emotion, 26*(3), 431–441. <https://doi.org/10.1080/02699931.2012.666502>
- Alley, T. R. (1988). Physiognomy and social perception. In T. R. Alley (Ed.), *Social and Applied Aspects of Perceiving Faces*. Hillsdale, NJ: Lawrence Erlbaum.
- Ames, D. R., Kammrath, L. K., Suppes, A., & Bolger, N. (2010). Not so fast: The (not-quite-complete) dissociation between accuracy and confidence in thin-slice impressions. *Personality and Social Psychology Bulletin, 26*(2), 264–277. <https://doi.org/10.1177/0146167209354519>
- An, J., & Weber, I. (2016). #greysanatomy vs. #yankees: Demographics and hashtag use on twitter. *Proceedings of the 10th International Conference on Weblogs and Social Media*, 523–526.
- Antonakis, J., & Dalgas, O. (2009). Predicting elections: Child's play! *Science, 323*, 1183. <https://doi.org/10.1126/science.1167748>
- Aristotle. (1936). Physiognomics. In T. E. Page, E. Capps, W. H. D. Rouse, A. Post, & E. H. Warmington (Eds.), *Minor works*. Cambridge, Massachusetts: Harvard University Press.
- Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258–290.
- Axt, J. R., Casola, G., & Nosek, B. A. (2018). Reducing social judgment biases may require identifying the potential source of bias. *Personality and Social Psychology Bulletin, 45*(8), 1232–1251. <https://doi.org/10.1177/0146167218814003>
- Axt, J. R., & Nosek, B. A. (2018). The Judgment Bias Task: A flexible method for assessing individual differences in social judgment biases. *Journal of Experimental Social Psychology, 76*, 337–355. <https://doi.org/10.1016/j.jesp.2018.02.011>
- Ayres, I., & Siegelman, P. (1995). Race and gender discrimination in bargaining for a new car. *The American Economic Review, 85*(3), 304–321.

- Bacharach, M., & Gambetta, D. (2001). Trust in signs. In K. S. Cook (Ed.), *Trust in Society* (pp. 148–184). New York: Russell Sage Foundation.
- Baert, S. (2018). Facebook profile picture appearance affects recruiters' first hiring decisions. *New Media and Society*, *20*(3), 1220–1239. <https://doi.org/10.1177/1461444816687294>
- Bainbridge, W. A., Isola, P., Blank, I., & Oliva, A. (2013). Establishing a database for studying human face photograph memory. In N. Miyake, D. Peebles, & R. P. Coopers (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1302–1307). Austin, TX: Cognitive Science Society.
- Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323–1334. <https://doi.org/10.1037/a0033872>
- Ballem, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(46), 17948–17953. <https://doi.org/10.1073/pnas.0705435104>
- Baratloo, A., Hosseini, M., Negida, A., & El Ashal, G. (2015). Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. *Emergency*, *3*(2), 48–49. Retrieved from <http://journals.sbmu.ac.ir/emergency/article/view/8154>
- Baumeister, R. F., & Monroe, A. E. (2014). Recent research on free will: Conceptualizations, beliefs, and processes. *Advances in Experimental Social Psychology*, *50*, 1–52. <https://doi.org/10.1016/B978-0-12-800284-1.00001-1>
- Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396–403. <https://doi.org/10.1111/j.1745-6916.2007.00051.x>
- Belot, M., Bhaskar, V., & van de Ven, J. (2010). Promises and cooperation: Evidence from a TV game show. *Journal of Economic Behavior and Organization*, *73*(3), 396–405. <https://doi.org/10.1016/j.jebo.2010.01.001>
- Belot, M., Bhaskar, V., & van den Ven, J. (2012). Can observers predict trustworthiness? *The Review of Economics and Statistics*, *94*(1), 246–259. <https://doi.org/10.1177/1745691612459060>

- Berg, J., Dickhaut, K., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142.
- Berggren, N., Jordahl, H., & Poutvaara, P. (2010). The looks of a winner: Beauty and electoral success. *Journal of Public Economics*, *94*(1–2), 8–15. <https://doi.org/10.1016/j.jpubeco.2009.11.002>
- Berggren, N., Jordahl, H., & Poutvaara, P. (2017). The right look: Conservative politicians look better and voters reward it. *Journal of Public Economics*, *146*, 79–86. <https://doi.org/10.1016/j.jpubeco.2016.12.008>
- Berry, D. S. (1990). Taking people at face value: Evidence for the kernel of truth hypothesis. *Social Cognition*, *8*(4), 343–361.
- Berry, D. S., & Zebrowitz-McArthur, L. A. (1988). What's in a face? Facial maturity and the attribution of legal responsibility. *Personality and Social Psychology Bulletin*, *14*(1), 23–33.
- Bettman, J. R., Johnson, E. J., & Payne, J. W. (1990). A componential analysis of cognitive effort in choice. *Organizational Behavior and Human Decision Processes*, *45*, 111–139.
- Biesanz, J. C., Human, L. J., Paquin, A.-C., Chan, M., Parisotto, K. L., Sarracino, J., & Gillis, R. L. (2011). Do we know when our impressions of others are valid? Evidence for realistic accuracy awareness in first impressions of personality. *Social Psychological and Personality Science*, *2*(5), 452–459. <https://doi.org/10.1177/1948550610397211>
- Bigoni, M., Bortolotti, S., Casari, M., Gambetta, D., & Pancotto, F. (2016). Amoral familism, social capital, or trust? The behavioural foundations of the Italian North–South divide. *Economic Journal*, *126*(594), 1318–1341. <https://doi.org/10.1111/ecoj.12292>
- Bindemann, M., Burton, A. M., Hooge, I. T. C., Jenkins, R., & de Haan, E. H. F. (2005). Faces retain attention. *Psychonomic Bulletin & Review*, *12*(6), 1048–1053. <https://doi.org/10.3758/BF03206442>
- Blair, I. V., Judd, C. M., Sadler, M. S., & Jenkins, C. (2002). The role of afrocentric features in person perception: Judging by features and categories. *Journal of Personality and Social Psychology*, *83*(1), 5–25. <https://doi.org/10.1037//0022-3514.83.1.5>
- Blair, I. V., Judd, C. M., & Chapleau, K. M. (2012). The influence of afrocentric facial features in criminal sentencing. *Psychological Science*, *15*(10), 674–679. <https://doi.org/10.1111/j.0956-7976.2004.00739.x>



- Blair, I. V., Judd, C. M., & Fallman, J. L. (2004). The automaticity of race and Afrocentric facial features in social judgments. *Journal of Personality and Social Psychology, 87*(6), 763–778.  
<https://doi.org/10.1037/0022-3514.87.6.763>
- Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2013). The modular nature of trustworthiness detection. *Journal of Experimental Psychology: General, 142*(1), 143–150.  
<https://doi.org/10.1037/a0028930>
- Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2015). Face-ism and kernels of truth in facial inferences. *Trends in Cognitive Sciences, 19*(8), 421–422. <https://doi.org/10.1016/j.tics.2015.05.002>
- Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2017). Can we detect cooperators by looking at their face? *Current Directions in Psychological Science, 26*(3), 276–281.  
<https://doi.org/10.1177/0963721417693352>
- Bóo, F. L., Rossi, M. A., & Urzúa, S. S. (2013). The labor market return to an attractive face: Evidence from a field experiment. *Economics Letters, 118*(1), 170–172.  
<https://doi.org/10.1016/j.econlet.2012.10.016>
- Borkenau, P., Brecke, S., Möttig, C., & Paelecke, M. (2009). Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality, 43*(4), 703–706.  
<https://doi.org/10.1016/j.jrp.2009.03.007>
- Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology, 62*(4), 645–657.
- Boyer, P., & Petersen, M. B. (2018). Folk-economic beliefs: An evolutionary cognitive model. *Behavioral and Brain Sciences, 41*, e158. <https://doi.org/10.1017/S0140525X17001960>
- Brambilla, M., Biella, M., & Freeman, J. B. (2018). The influence of visual context on the evaluation of facial trustworthiness. *Journal of Experimental Social Psychology, 78*, 34–42.  
<https://doi.org/10.1016/j.jesp.2018.04.011>
- Brambilla, M., Carraro, L., Castelli, L., & Sacchi, S. (2019). Changing impressions: Moral character dominates impression updating. *Journal of Experimental Social Psychology, 82*, 64–73.  
<https://doi.org/S0022103118303184>

- Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology, 41*(2), 135–143. <https://doi.org/10.1002/ejsp.744>
- Bruce, V., & Young, A. (2012). *Face perception*. London; New York: Psychology Press.
- Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley: Univer. California Press.
- Buss, D. M. (1989). Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioural and Brain Sciences, 12*(1989), 1–14. <https://doi.org/10.1017/S0140525X00023992>
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. <https://doi.org/10.1038/nrn3475>
- Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116–131.
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., ... Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2*, 637–644. <https://doi.org/10.1038/s41562-018-0399-z>
- Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology, 67*(2), 319–333. <https://doi.org/10.1037/0022-3514.67.2.319>
- Castelli, L., Carraro, L., Ghitti, C., & Pastore, M. (2009). The effects of perceived competence and sociability on electoral outcomes. *Journal of Experimental Social Psychology, 45*(5), 1152–1155. <https://doi.org/10.1016/j.jesp.2009.06.018>
- Caulfield, F., Ewing, L., Bank, S., & Rhodes, G. (2016). Judging trustworthiness from faces: Emotion cues modulate trustworthiness judgments in young children. *British Journal of Psychology, 107*(3), 503–518. <https://doi.org/10.1111/bjop.12156>

- Centorrino, S., Djemai, E., Hopfensitz, A., Milinski, M., & Seabright, P. (2015). Honest signaling in trust interactions: Smiles rated as genuine induce trust and signal higher earning opportunities. *Evolution and Human Behavior, 36*(1), 8–16.  
<https://doi.org/10.1016/j.evolhumbehav.2014.08.001>
- Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*(5), 752–766.
- Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of sourcecredibility, argument ambiguity, and task importance on attitude judgment. *Journal of Personality and Social Psychology, 66*(3), 460–473.
- Chan, M.-P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science, 28*(11), 1531–1546.  
<https://doi.org/10.1177/09567976177114579>
- Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87–105.  
<https://doi.org/10.1016/j.cogpsych.2010.03.001>
- 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, 55*(7), 1400–1413. <https://doi.org/10.1037/dev0000734>
- Chen, E. E., & Wojcik, S. P. (2016). A practical guide to Big Data research in psychology. *Psychological Methods, 21*(4), 458–474.  
<https://doi.org/http://dx.doi.org/10.1037/met0000111>
- Chen, F. F., Jing, Y., & Lee, J. M. (2012). “I” value competence but “we” value social competence: The moderating role of voters' individualistic and collectivistic orientation in political elections. *Journal of Experimental Social Psychology, 48*(6), 1350–1355.  
<https://doi.org/10.1016/j.jesp.2012.07.006>
- Chen, F. F., Jing, Y., & Lee, J. M. (2014). The looks of a leader: Competent and trustworthy, but not dominant. *Journal of Experimental Social Psychology, 51*, 27–33.  
<https://doi.org/10.1016/j.jesp.2013.10.008>

- Chen, F. F., Jing, Y., Lee, J. M., & Bai, L. (2016). Culture matters: The looks of a leader are not all the same. *Social Psychological and Personality Science*, 7(6), 570–578.  
<https://doi.org/10.1177/1948550616644962>
- Chiao, J. Y., Bowman, N. E., & Gill, H. (2008). The political gender gap: Gender bias in facial inferences that predict voting behavior. *PLoS ONE*, 3(10). <https://doi.org/10.1371/journal.pone.0003666>
- Chiu, C., Hong, Y., & Dweck, C. S. (1997). Lay dispositionism and implicit theories of personality. *Journal of Personality and Social Psychology*, 73(1), 19–30. <https://doi.org/10.1037/0022-3514.73.1.19>
- Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, 88, 2–9. <https://doi.org/10.1111/j.1475-4932.2012.00809.x>
- Clark, H. H., & Haviland, S. E. (1977). Comprehension and the given-new contract. In R. O. Freedle (Ed.), *Discourse production and comprehension* (pp. 1–40). <https://doi.org/10.2307/1421524>
- Cleeton, G. U., & Knight, F. B. (1924). Validity of character judgments based on external criteria. *Journal of Applied Psychology*, 8(2), 215–231.
- Cogsdill, E. J., & Banaji, M. R. (2015). Face-trait inferences show robust child–adult agreement: Evidence from three types of faces. *Journal of Experimental Social Psychology*, 60, 150–156.  
<https://doi.org/10.1016/j.jesp.2015.05.007>
- Collins, A. F. (1999). The enduring appeal of physiognomy: Physical appearance as a sign of temperament, character, and intelligence. *History of Psychology*, 2(4), 251–276.  
<https://doi.org/10.1037/1093-4510.2.4.251>
- Costa-Gomes, M. A., Huck, S., & Weizsäcker, G. (2014). Beliefs and actions in the trust game: Creating instrumental variables to estimate the causal effect. *Games and Economic Behavior*, 88, 298–309. <https://doi.org/10.1016/j.geb.2014.10.006>
- Cox, G. W., & Katz, J. N. (1996). Why did the incumbency advantage in U.S. House elections grow? *American Journal of Political Science*, 40(2), 478–497.

- Cox, W. T. L., Devine, P. G., Bischmann, A. A., Hyde, J. S., Cox, W. T. L., Devine, P. G., ... Hyde, J. S. (2015). Inferences about sexual orientation: The roles of stereotypes, faces, and the Gaydar myth. *The Journal of Sex Research, 53*(2), 157–171.  
<https://doi.org/10.1080/00224499.2015.1015714>
- Crivelli, C., & Fridlund, A. J. (2018). Facial displays are tools for social influence. *Trends in Cognitive Sciences, 22*(5), 388–399.  
<https://doi.org/10.1016/j.tics.2018.02.006>
- Cunningham, M. R., Roberts, A. R., Barbee, A. P., & Druen, P. B. (1995). “Their ideas of beauty are, on the whole, the same as ours”: Consistency and variability in the cross-cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology, 68*(2), 261–279.
- Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of Personality and Social Psychology, 90*(5), 784–803.  
<https://doi.org/10.1037/0022-3514.90.5.784>
- Darai, D., & Grätz, S. (2013). *Attraction and cooperative behavior*. Retrieved from  
<http://www.neweconomists.org/files/Attraction.pdf>
- Darwin, C. (1887). *The autobiography of Charles Darwin*. Barnes & Noble Publishing.
- Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology, 31*, 169–193.  
<https://doi.org/10.1146/annurev.ps.31.020180.001125>
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668–1674.  
<https://doi.org/10.1126/science.2648573>
- De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE, 6*(1).  
<https://doi.org/10.1371/journal.pone.0015954>
- De Neys, W., Hopfensitz, A., & Bonnefon, J.-F. (2013). Low second-to-fourth digit ratio predicts indiscriminate social suspicion, not improved trustworthiness detection. *Biology Letters, 9*(2), 20130037. <https://doi.org/10.1098/rsbl.2013.0037>
- De Neys, W., Hopfensitz, A., & Bonnefon, J. F. (2017). Split-second trustworthiness detection from faces in an economic game. *Experimental Psychology, 64*, 231–239.  
<https://doi.org/10.1027/1618-3169/a000367>

- Deaner, R. O., Goetz, S. M. M., Shattuck, K., & Schnotala, T. (2012). Body weight, not facial width-to-height ratio, predicts aggression in pro hockey players. *Journal of Research in Personality, 46*(2), 235–238. <https://doi.org/10.1016/j.jrp.2012.01.005>
- DeBruine, L. M., & Jones, B. C. (2017). *Face Research Lab London Set*. <https://doi.org/10.6084/m9.figshare.5047666.v3>
- Del Monte, A., & Papagni, E. (2007). The determinants of corruption in Italy: Regional panel data analysis. *European Journal of Political Economy, 23*(2), 379–396. <https://doi.org/10.1016/j.ejpoleco.2006.03.004>
- Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch’s t-test instead of Student’s t-test. *International Review of Social Psychology, 30*(1), 92–101. <https://doi.org/10.5334/irsp.82>
- Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5–18.
- Devine, P. G., Forscher, P. S., & Austin, A. J. (2013). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology, 48*(6), 1267–1278. <https://doi.org/10.1016/j.jesp.2012.06.003>
- Dimov, C. M., & Link, D. (2017). Do people order cues by retrieval fluency when making probabilistic inferences? *Journal of Behavioral Decision Making*. <https://doi.org/10.1002/bdm.2002>
- Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24*(3), 285–290. <https://doi.org/10.1177/0956797610388048>
- Doleac, J. L., & Stein, L. C. D. (2013). The visible hand: Race and online market outcomes. *The Economic Journal, 123*(572). <https://doi.org/10.1111/econj.12082>
- Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science, 11*(4), 315–319.
- Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*(1), 62–68. <https://doi.org/10.1037/0022-3514.82.1.62>
- Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. *Review of Financial Studies, 25*(8), 2455–2483. <https://doi.org/10.1093/rfs/hhs071>

- 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>
- Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior. *American Psychologist*, *54*(6), 408–423. <https://doi.org/10.1037/0003-066X.54.6.408>
- Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy. *Psychological Science*, *17*(5), 383–386. <https://doi.org/10.1111/j.1467-9280.2006.01716.x>
- Eckel, C. C., & Grossman, P. J. (1998). Are women less selfish than men? Evidence from dictator experiments. *The Economic Journal*, *108*(448), 726–735.
- Eckel, C. C., & Petrie, R. (2011). Face value. *The American Economic Review*, *101*(4), 1497–1513. <https://doi.org/10.1257/aer.101.4.1497>
- Edelman, B. (2016). *Response to Airbnb's report on discrimination*. Retrieved from <http://www.benedelman.org/news-091916/>
- Edelman, B., & Luca, M. (2014). Digital Discrimination: The Case of Airbnb.com. In *Harvard Business School NOM Unit Working Paper No. 14-054*. Retrieved from <https://ssrn.com/abstract=2377353>
- Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, *9*(2), 1–22. <https://doi.org/10.1257/app.20160213>
- Efferson, C., & Vogt, S. (2013). Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports*, *3*, 1047. <https://doi.org/10.1038/srep01047>
- Einav, L., Farronato, C., & Levin, J. (2016). Peer-to-peer markets. *Annual Review of Economics*, *8*(1), 615–635. <https://doi.org/10.1146/annurev-economics-080315-015334>
- Elliot, A. J., & Thrash, T. M. (2002). Approach-avoidance motivation in personality: Approach and avoidance temperaments and goals. *Journal of Personality and Social Psychology*, *82*(5), 804–818. <https://doi.org/10.1037/0022-3514.82.5.804>
- Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience*, *19*(9), 1508–1519. <https://doi.org/10.1162/jocn.2007.19.9.1508>

- Epitropaki, O., & Martin, R. (2004). Implicit leadership theories in applied settings: Factor structure, generalizability, and stability over time. *Journal of Applied Psychology, 89*(2), 293–310. <https://doi.org/10.1037/0021-9010.89.2.293>
- Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of Personality and Social Psychology, 71*(2), 390–405. <https://doi.org/10.1037/0022-3514.71.2.390>
- Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism Management, 55*, 62–73. <https://doi.org/10.1016/j.tourman.2016.01.013>
- Evans, A. M., Dillon, K. D., & Rand, D. G. (2015). Fast but not intuitive, slow but not reflective: Decision conflict drives reaction times in social dilemmas. *Journal of Experimental Psychology: General, 144*(5), 951–966. <https://doi.org/10.1037/xge0000107>
- Evans, A. M., & Krueger, J. I. (2011). Elements of trust: Risk and perspective-taking. *Journal of Experimental Social Psychology, 47*(1), 171–177. <https://doi.org/10.1016/j.jesp.2010.08.007>
- Evans, A. M., & Krueger, J. I. (2014). Outcomes and expectations in dilemmas of trust. *Judgment and Decision Making, 9*(2), 90–103.
- Evans, A. M., & Krueger, J. I. (2016). Bounded prospection in dilemmas of trust and reciprocity. *Reviews of General Psychology, 20*(1), 17–28. <https://doi.org/10.1007/s13398-014-0173-7.2>
- Evans, A. M., & Rand, D. G. (2019). Cooperation and decision time. *Current Opinion in Psychology, 26*, 67–71. <https://doi.org/10.1016/j.copsyc.2018.05.007>
- Evans, A. M., & van de Calseyde, P. P. F. M. (2017). The effects of observed decision time on expectations of extremity and cooperation. *Journal of Experimental Social Psychology, 68*, 50–59. <https://doi.org/10.1016/j.jesp.2016.05.009>
- Ewing, L., Caulfield, F., Read, A., & Rhodes, G. (2014). Perceived trustworthiness of faces drives trust behaviour in children. *Developmental Science, 2*, 327–334. <https://doi.org/10.1111/desc.12218>
- Ewing, L., Caulfield, F., Read, A., & Rhodes, G. (2015). Appearance-based trust behaviour is reduced in children with autism spectrum disorder. *Autism, 19*(8), 1002–1009. <https://doi.org/10.1177/1362361314559431>



- Fagerstrøm, A., Pawar, S., Sigurdsson, V., Foxall, G. R., & Yani-de-Soriano, M. (2017). That personal profile image might jeopardize your rental opportunity! On the relative impact of the seller's facial expressions upon buying behavior on Airbnb™. *Computers in Human Behavior*, *72*, 123–131. <https://doi.org/10.1016/j.chb.2017.02.029>
- Fang, X., van Kleef, G. A., & Sauter, D. A. (2018). Person perception from changing emotional expressions: primacy, recency, or averaging effect? *Cognition and Emotion*, *32*(8), 1597–1610. <https://doi.org/10.1080/02699931.2018.1432476>
- Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is “special” about face processing? *Psychological Review*, *105*(3), 482–498.
- Farroni, T., Johnson, M. H., Menon, E., Zulian, L., Faraguna, D., & Csibra, G. (2005). Newborns' preference for face-relevant stimuli: Effects of contrast polarity. *Proceedings of the National Academy of Sciences*, *102*(47), 17245–17250. <https://doi.org/10.1073/pnas.0502205102>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Feingold, A. (1990). Gender differences in effects of physical attractiveness on romantic attraction: A comparison across five research paradigms. *Journal of Personality and Social Psychology*, *59*(5), 981–993. <https://doi.org/10.1037/0022-3514.59.5.981>
- Feliciano, C., Robnett, B., & Komaie, G. (2009). Gendered racial exclusion among white internet daters. *Social Science Research*, *38*(1), 39–54. <https://doi.org/10.1016/j.ssresearch.2008.09.004>
- Ferguson, H. S., Owen, A., Hahn, A. C., Torrance, J., DeBruine, L. M., & Jones, B. C. (2019). Context-specific effects of facial dominance and trustworthiness on hypothetical leadership decisions. *PLoS ONE*, *14*(7), e0214261. <https://doi.org/10.1371/journal.pone.0214261>
- Fiorino, N., Galli, E., & Petrarca, I. (2012). Corruption and growth: Evidence from Italian regions. *European Journal of Government and Economics*, *1*(2), 126–144.

- Fiske, A. P., Haslam, N., & Fiske, S. T. (1991). Confusing one person with another: What errors reveal about the elementary forms of social relations. *Journal of Personality and Social Psychology*, *60*(5), 656–674.
- Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902.  
<https://doi.org/10.1037//0022-3514.82.6.878>
- Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change in implicit bias. *Journal of Personality and Social Psychology*, *117*(3), 522–559. <https://doi.org/10.1037/pspa0000160>
- Fraley, R. C., & Vazire, S. (2014). The N-Pact Factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE*, *9*(10), e109019.  
<https://doi.org/10.1371/journal.pone.0109019>
- Frankenhuis, W., & Nettle, D. (2018). Open Science is liberating and can foster creativity. *Perspectives on Psychological Science*, *13*(4), 439–447. <https://doi.org/10.1177/1745691618767878>
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.  
<https://doi.org/10.1257/089533005775196732>
- Freeman, J. B., & Ambady, N. (2010). MouseTracker: software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–241.  
<https://doi.org/10.3758/BRM.42.1.226>
- Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247–279.  
<https://doi.org/10.1037/a0022327>
- Freeman, J. B., & Johnson, K. L. (2016). More than meets the eye: Split-second social perception. *Trends in Cognitive Sciences*, *20*(5), 362–374. <https://doi.org/10.1016/j.tics.2016.03.003>
- Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, *21*(3), 177–182.  
<https://doi.org/10.1177/0963721412445309>

- Gheorghiu, A. I., Callan, M. J., & Skylark, W. J. (2017). Facial appearance affects science communication. *Proceedings of the National Academy of Sciences*, *114*(23), 5970–5975. <https://doi.org/10.1073/pnas.1620542114>
- Gibbs, C., Guttentag, D., Gretzel, U., Morton, J., & Goodwill, A. (2018). Pricing in the sharing economy: A hedonic pricing model applied to Airbnb listings. *Journal of Travel and Tourism Marketing*, *35*(1), 46–56. <https://doi.org/10.1080/10548408.2017.1308292>
- Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. <https://doi.org/10.1037/0033-295X.103.4.650>
- Gigerenzer, G., Hertwig, R., & Pachur, T. (2011). *Heuristics: The foundations of adaptive behavior*. New York, NY: Oxford University Press.
- Gigerenzer, G., Martignon, L., Hoffrage, U., Rieskamp, J., Czerlinski, J., & Goldstein, D. G. (2008). One-reason decision making. In C. R. Plott & V. L. Smith (Eds.), *Handbook of Experimental Economics* (Vol. 1, pp. 1004–1017). Amsterdam: North-Holland.
- Gladstone, E., & O'Connor, K. M. (2014). A counterpart's feminine face signals cooperativeness and encourages negotiators to compete. *Organizational Behavior and Human Decision Processes*, *125*(1), 18–25. <https://doi.org/10.1016/j.obhdp.2014.05.001>
- Gomulya, D., Wong, E. M., Ormiston, M. E., & Boeker, W. (2017). The role of facial appearance on CEO selection after firm misconduct. *Journal of Applied Psychology*, *102*(4), 617–635. <https://doi.org/http://dx.doi.org/10.1037/apl0000172>
- Gonzalez, L., & Loureiro, Y. K. (2014). When can a photo increase credit? The impact of lender and borrower profiles on online peer-to-peer loans. *Journal of Behavioral and Experimental Finance*, *2*, 44–58. <https://doi.org/10.1016/j.jbef.2014.04.002>
- Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science*, *24*(1), 38–44. <https://doi.org/10.1177/0963721414550709>
- Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. <https://doi.org/10.1037/a0034726>

- Gordon, R. A., & Arvey, R. D. (2004). Age bias in laboratory and field settings: A meta-analytic investigation. *Journal of Applied Social Psychology, 34*(3), 468–492. <https://doi.org/10.1111/j.1559-1816.2004.tb02557.x>
- Graham, J. R., Harvey, C. R., & Puri, M. (2017). A corporate beauty contest. *Management Science, 63*(9), 3044–3056. <https://doi.org/10.1287/mnsc.2016.2484>
- Green, P., & Macleod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*, 493–498. <https://doi.org/10.1111/2041-210X.12504>
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27. <https://doi.org/10.1037/0033-295X.102.1.4>
- Greenwald, A. G., McGhee, D. E., Jordan, L. K., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480.
- Gunnell, J. J., & Ceci, S. J. (2010). When emotionality trumps reason: A study of individual processing style and juror bias. *Behavioral Sciences & The Law, 28*(2), 211–223. <https://doi.org/10.1002/bsl>
- Gutta, S., Huang, J. R. J., Jonathon, P., & Wechsler, H. (2000). Mixture of experts for classification of gender, ethnic origin, and pose of human faces. *IEEE Transactions on Neural Networks, 11*(4), 948–960. <https://doi.org/10.1109/72.857774>
- Guttentag, D. (2013). Airbnb: Disruptive innovation and the rise of an informal tourism accommodation sector. *Current Issues in Tourism, 37*–41. <https://doi.org/10.1080/13683500.2013.827159>
- Hamermesh, D. S. (2011). *Beauty pays: Why attractive people are more successful*. <https://doi.org/10.1017/CBO9781107415324.004>
- Hamermesh, D. S., & Biddle, J. E. (1994). Beauty and the labor market. *The American Economic Review, 84*(5), 1174–1194. <https://doi.org/10.1016/B978-0-08-097086-8.94015-7>
- Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review, 71*(6), 438–456. <https://doi.org/10.1037/h0040736>

- Han, C., Wang, H., Hahn, A. C., Fisher, C. I., Fasolt, V., Morrison, D. K., ... Jones, B. C. (2017). *Cultural differences in preferences for facial coloration*. Retrieved from <https://psyarxiv.com/8zae5/>
- Haslam, N., Bastian, B., & Bissett, M. (2004). Essentialist beliefs about personality and their implications. *Personality and Social Psychology Bulletin, 30*, 1661–1673. <https://doi.org/10.1177/0146167204271182>
- Hassin, R., & Trope, Y. (2000). Facing faces: Studies on the cognitive aspects of physiognomy. *Journal of Personality and Social Psychology, 78*(5), 837–852. <https://doi.org/10.1037//0022-3514.78.5.837>
- Hehman, E., Carpinella, C. M., Johnson, K. L., Leitner, J. B., & Freeman, J. B. (2014). Early processing of gendered facial cues predicts the electoral success of female politicians. *Social Psychological and Personality Science, 5*(7), 815–824. <https://doi.org/10.1177/1948550614534701>
- Hehman, E., Leitner, J. B., Deegan, M. P., & Gaertner, S. L. (2015). Picking teams: When dominant facial structure is preferred. *Journal of Experimental Social Psychology, 59*, 61–65. <https://doi.org/10.1016/j.jesp.2015.03.007>
- Hehman, E., Leitner, J. B., & Gaertner, S. L. (2013). Enhancing static facial features increases intimidation. *Journal of Experimental Social Psychology, 49*(4), 747–754. <https://doi.org/10.1016/j.jesp.2013.02.015>
- Hehman, E., Stolier, R. M., Freeman, J. B., Flake, J. K., & Xie, S. Y. (2019). Toward a comprehensive model of face impressions: What we know, what we do not, and paths forward. *Social and Personality Psychology Compass, 13*(2), 1–16. <https://doi.org/10.1111/spc3.12431>
- Hehman, E., Sutherland, C. A. M., Flake, J. K., & Slepian, M. L. (2017). The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology, 113*(4), 513–529. <https://doi.org/10.1037/pspa0000090>
- Hehman, E., Xie, S. Y., Ofosu, E. K., & Nespoli, G. A. (2018). *Assessing the point at which averages are stable: A tool illustrated in the context of person perception*. Retrieved from <https://psyarxiv.com/2n6jq/>

- Hershberger, P. J., Markert, R. J., Part, H. M., Cohen, S. M., & Finger, W. W. (1997). Understanding and addressing cognitive bias in medical education. *Advances in Health Sciences Education, 1*, 221–226.  
<https://doi.org/10.1007/BF00162919>
- Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(5), 1191–1206.  
<https://doi.org/10.1037/a0013025>
- Heyes, A., & List, J. A. (2016). Supply and demand for discrimination: Strategic revelation of own characteristics in a trust game. *American Economic Review, 106*(5), 319–323.  
<https://doi.org/10.1257/aer.p20161011>
- Hollander, E. P., & Julian, J. W. (1969). Contemporary trends in the analysis of leadership processes. *Psychological Bulletin, 71*(5), 387–397.
- Hönekopp, J. (2006). Once more: is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance, 32*(2), 199–209.  
<https://doi.org/10.1037/0096-1523.32.2.199>
- Huang, W., Weber, I., & Vieweg, S. (2014). Inferring nationalities of Twitter users and studying inter-national linking. *Proceedings of the 25th ACM Conference on Hypertext and Social Media - HT '14*, 237–242. <https://doi.org/10.1145/2631775.2631825>
- Hugenberg, K., & Wilson, J. P. (2013). Faces are central to social cognition. In D. E. Carlston (Ed.), *Handbook of Social Cognition* (pp. 167–193).  
<https://doi.org/10.1093/oxfordhb/9780199730018.013.0009>
- Hung, S.-M., Nieh, C.-H., & Hsieh, P.-J. (2016). Unconscious processing of facial attractiveness: Invisible attractive faces orient visual attention. *Scientific Reports, 6*, 37117.  
<https://doi.org/10.1038/srep37117>
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), 0696–0701.  
<https://doi.org/10.1371/journal.pmed.0020124>

- Jäckle, S., Metz, T., Wenzelburger, G., & König, P. D. (2019). A catwalk to congress? Appearance-based effects in the elections to the U.S. House of Representatives 2016. *American Politics Research*. <https://doi.org/10.1177/1532673X19875710>
- Jadidi, M., Karimi, F., Lietz, H., & Wagner, C. (2017). *Gender disparities in science? Dropout, productivity, collaborations and success of male and female computer scientists*. Retrieved from <https://arxiv.org/abs/1704.05801>
- Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2019a). 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>
- Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2019b). *Who judges a book by its cover? The prevalence, structure, and correlates of lay beliefs in physiognomy*. Retrieved from <https://psyarxiv.com/8dq4x/>
- Jaeger, B., Evans, A. M., & van Beest, I. (2019). *Facial appearance and electoral success: Are trustworthy-looking politicians more successful in corrupt regions?* Retrieved from <https://psyarxiv.com/btcxm/>
- Jaeger, B., Slegers, W. W. A., & Evans, A. M. (2019). Automated classification of demographics from face images: A tutorial and validation. *Social and Personality Psychology Compass*.
- Jaeger, B., Slegers, W. W. A., Evans, A. M., Stel, M., & van Beest, I. (2019). The effects of facial attractiveness and trustworthiness in online peer-to-peer markets. *Journal of Economic Psychology*, *75 Part A*, 102125. <https://doi.org/10.1016/j.joep.2018.11.004>
- Jaeger, B., Todorov, A. T., Evans, A. M., & van Beest, I. (2019). *Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions*. Retrieved from <https://psyarxiv.com/a8w2d/>
- Jaeger, B., Wagemans, F. M. A., Evans, A. M., & van Beest, I. (2018). Effects of facial skin smoothness and blemishes on trait impressions. *Perception*, *47*(6), 608–625. <https://doi.org/10.1177/0301006618767258>
- Jain, A. K. (2001). Corruption: A review. *Journal of Economic Surveys*, *15*(1), 71–121. <https://doi.org/10.1111/1467-6419.00133>
- Jarvis, W. B. G., & Petty, R. E. (1996). The need to evaluate. *Journal of Personality and Social Psychology*, *70*(1), 172–194.

- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532.  
<https://doi.org/10.1177/0956797611430953>
- Johnson, E. J., Payne, J. W., Schkade, D. A., & Bettman, J. R. (1989). *Monitoring information processing and decisions: The Mouselab system.*
- Jones, A. L., & Jaeger, B. (2019). Biological bases of beauty revisited: The effect of symmetry, averageness, and sexual dimorphism on female facial attractiveness. *Symmetry, 11*(2).  
<https://doi.org/10.3390/sym11020279>
- Jones, A. L., Kramer, R. S. S., & Ward, R. (2012). Signals of personality and health: The contributions of facial shape, skin texture, and viewing angle. *Journal of Experimental Psychology: Human Perception and Performance, 38*(6), 1353–1361.  
<https://doi.org/10.1037/a0027078>
- Jones, B. C., DeBruine, L. M., Flake, J. K., Aczel, B., Adamkovic, M., Alaei, R., ... Chartier, C. R. (2019). *To which world regions does the valence-dominance model of social perception apply?* Retrieved from <https://psyarxiv.com/n26dy/>
- Jones, C. S., & Kaplan, M. F. (2003). The effects of racially stereotypical crimes on juror decision-making and information-processing strategies. *Basic and Applied Social Psychology, 25*(1), 1–13.  
<https://doi.org/10.1207/S15324834BASP2501>
- Kakar, V., Franco, J., Voelz, J., & Wu, J. (2016). *The visible host: Does race guide Airbnb rental rates in San Francisco?* Retrieved from <https://mpr.aub.uni-muenchen.de/78275/>
- Kelley, H. H., Holmes, J. G., Kerr, Norbert, L., Reis, H. T., Rusbult, C. E., & Van Lange, P. A. M. (2003). *An Atlas of Interpersonal Situations.* Cambridge University Press.
- Kenny, D. A., & West, T. V. (2008). Zero acquaintance: Definitions, statistical model, findings, and process. In N. Ambady & J. J. Skowronski (Eds.), *First impressions* (pp. 129–146). New York: Guilford Press.
- Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., ... Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology, 14*(5), 1–15.  
<https://doi.org/10.1371/journal.pbio.1002456>



- King, A., & Leigh, A. (2009). Beautiful politicians. *Kyklos*, 62(4), 579–593. [https://doi.org/10.1016/0014-2921\(83\)90040-5](https://doi.org/10.1016/0014-2921(83)90040-5)
- Klapper, A., Dotsch, R., van Rooij, I., & Wigboldus, D. H. J. (2016). Do we spontaneously form stable trustworthiness impressions from facial appearance? *Journal of Personality and Social Psychology*, 111(5), 655–664. <https://doi.org/10.1037/pspa0000062>
- Knutson, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior*, 20(3), 165–182. <https://doi.org/10.1007/bf02281954>
- Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, 139(4), 665–682. <https://doi.org/10.1037/a0020198>
- Kosinski, M. (2017). Facial width does not predict self-reported behavioral tendencies. *Psychological Science*, 28(11), 1675–1682. <https://doi.org/10.1177/0956797617716929>
- Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences. *American Psychologist*, 70(6), 543–556. <https://doi.org/10.1037/a0039210>
- Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining Big Data to extract patterns and predict real-life outcomes. *Psychological Methods*, 21(4), 493–506. <https://doi.org/10.1037/met0000105>
- Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, 6, 7455. <https://doi.org/10.1038/ncomms8455>
- Kramer, R. S. S., & Ward, R. (2010). Internal facial features are signals of personality and health. *Quarterly Journal of Experimental Psychology*, 63(11), 2273–2287. <https://doi.org/10.1080/17470211003770912>
- Kring, A. M., Smith, D. A., & Neale, J. M. (1994). Individual differences in dispositional expressiveness: Development and validation of the Emotional Expressivity Scale. *Journal of Personality and Social Psychology*, 66(5), 934–949. <https://doi.org/10.1037/0022-3514.66.5.934>

- Krumhuber, E. G., Manstead, A. S. R. R., Cosker, D., Marshall, D., Rosin, P. L., & Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion, 7*(4), 730–735. <https://doi.org/10.1037/1528-3542.7.4.730>
- Kuklinski, J. H., & Quirk, P. J. (2000). Reconsidering the rational public: Cognition, heuristics, and mass opinion. In A. Lupia, M. D. McCubbins, & S. L. Popkin (Eds.), *Elements of reason: Cognition, choice, and the bounds of rationality* (pp. 153–182). Cambridge: Cambridge University Press.
- Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Joy-Gaba, J. A., Ho, A. K., ... Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Social Psychology, 143*(4), 1765–1785. <https://doi.org/10.1037/a0036260>
- Landers, R. N., Brusso, R., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of Big Data from the internet for use in psychological research. *Psychological Methods, 21*(4), 475–492. <https://doi.org/10.1037/a0033269>
- Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., ... Uhlmann, E. L. (n.d.). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*. Retrieved from <http://home.uchicago.edu/bartels/papers/Landy-et-al.-2020-PsychologicalBulletin.pdf>
- Landy, J. F., Piazza, J., & Goodwin, G. P. (2016). When it's bad to be friendly and smart: The desirability of sociability and competence depends on morality. *Personality and Social Psychology Bulletin, 42*(9), 1272–1290. <https://doi.org/10.1177/0146167216655984>
- Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hauam, M., Smoot, M., ... Swann, W. B. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin, 126*(3), 390–423. <https://doi.org/10.1037//0033-2909.126.3.390>
- Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion, 24*(8), 1377–1388. <https://doi.org/10.1080/02699930903485076>

- Laustsen, L., & Petersen, M. B. (2015). Does a competent leader make a good friend? Conflict, ideology and the psychologies of friendship and followership. *Evolution and Human Behavior, 36*(4), 286–293. <https://doi.org/10.1016/j.evolhumbehav.2015.01.001>
- Lavater, J. C. (1775). *Essays on Physiognomy: Designed to Promote the Knowledge and the Love of Mankind*. London: William Tegg and Co.
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... Van Alstyne, M. (2009). Life in the network: The coming age of computational social science. *Science, 323*(5915), 721–723. <https://doi.org/10.1126/science.1167742.Life>
- Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology, 93*(2), 234–249. <https://doi.org/10.1037/0022-3514.93.2.234>
- Leleu, A., Dzhelyova, M., Rossion, B., Brochard, R., Durand, K., Schaal, B., & Baudouin, J. Y. (2018). Tuning functions for automatic detection of brief changes of facial expression in the human brain. *NeuroImage, 179*, 235–251. <https://doi.org/10.1016/j.neuroimage.2018.06.048>
- Lenz, G. S., & Lawson, C. (2011). Looking the part: Television leads less informed citizens to vote based on candidates' appearance. *American Journal of Political Science, 55*(3), 574–589. <https://doi.org/10.1111/j.1540-5907.2011.00511.x>
- Leopold, D. A., & Rhodes, G. (2010). A comparative view of face perception. *Journal of Comparative Psychology, 124*(3), 233–251. <https://doi.org/10.1037/a0019460>
- Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 34–42. <https://doi.org/10.1109/CVPRW.2015.7301352>
- Levy, S. R., Stroessner, S. J., & Dweck, C. S. (1998). Stereotype formation and endorsement: The role of implicit theories. *Journal of Personality and Social Psychology, 74*(6), 1421–1536. <https://doi.org/10.1037/0022-3514.74.6.1421>
- 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>

- Liggett, J. (1974). *The human face*. New York: Stein & Day.
- Lin, C., Adolphs, R., & Alvarez, R. M. (2017). Cultural effects on the association between election outcomes and face-based trait inferences. *PLoS ONE*, *12*(7), e0180837.  
<https://doi.org/10.1371/journal.pone.0180837>
- Lin, C., Adolphs, R., & Alvarez, R. M. (2018). Inferring whether officials are corruptible from looking at their faces. *Psychological Science*, *29*(11), 1807–1823.  
<https://doi.org/10.1177/0956797618788882>
- Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551–556.  
<https://doi.org/10.1016/j.jesp.2009.12.019>
- Ling, L., Luo, D., & She, G. (2019). Judging a book by its cover: The influence of physical attractiveness on the promotion of regional leaders. *Journal of Economic Behavior and Organization*, (158), 1–14. <https://doi.org/10.1016/j.jebo.2019.01.005>
- Linhartova, V., & Pultarova, M. (2015). Cross-regional comparison of corruption. *4th International Conference on Economics, Political and Law Science*, 118–124. Retrieved from <http://www.wseas.us/e-library/conferences/2015/Rome/EPLS/EPLS-14.pdf>
- Lipkus, I. M. (1991). The construction and preliminary validation of a global Belief in a Just World Scale and the exploratory analysis of the multidimensional Belief in a Just World Scale. *Personality and Individual Differences*, *12*(11), 1171–1178.
- Little, A. C. (2014). Facial appearance and leader choice in different contexts: Evidence for task contingent selection based on implicit and learned face-behaviour/face-ability associations. *Leadership Quarterly*, *25*(5), 865–874.  
<https://doi.org/10.1016/j.leaqua.2014.04.002>
- Little, A. C., Burriss, R. P., Jones, B. C., & Roberts, S. C. (2007). Facial appearance affects voting decisions. *Evolution and Human Behavior*, *28*(1), 18–27.  
<https://doi.org/10.1016/j.evolhumbehav.2006.09.002>
- Little, A. C., & Perrett, D. I. (2007). Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology*, *98*(1), 111–126.  
<https://doi.org/10.1348/000712606X109648>

- Little, A. C., Roberts, S. C., Jones, B. C., & DeBruine, L. M. (2012). The perception of attractiveness and trustworthiness in male faces affects hypothetical voting decisions differently in wartime and peacetime scenarios. *The Quarterly Journal of Experimental Psychology*, *65*(10), 2018–2032. <https://doi.org/10.1080/17470218.2012.677048>
- Lu, X., & Jain, A. (2004). Ethnicity identification from face images. *Proceedings of SPIE Conference on Biometric Technology for Human Identification*, *5404*, 114–123. <https://doi.org/10.1117/12.542847>
- 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>
- Ma, X., Hancock, J. T., Mingjie, K. L., & Naaman, M. (2017). Self-disclosure and perceived trustworthiness of Airbnb host profiles. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*, 2397–2409. <https://doi.org/10.1145/2998181.2998269>
- Maestriperi, D., Henry, A., & Nickels, N. (2017). Explaining financial and prosocial biases in favor of attractive people: Interdisciplinary perspectives from economics, social psychology, and evolutionary psychology. *Behavioral and Brain Sciences*, *40*, e19. <https://doi.org/10.1017/S0140525X16000340>
- Malpezzi, S. (2008). Hedonic pricing models: A selective and applied review. In T. O'Sullivan & K. Gibb (Eds.), *Housing Economics and Public Policy*. Sage Publications, Inc.
- Maner, J. K. (2016). Into the wild: Field research can increase both replicability and real-world impact. *Journal of Experimental Social Psychology*, *66*, 100–106. <https://doi.org/10.1016/j.jesp.2015.09.018>
- Marsh, A. A. (2005). Why do fear and anger look the way they do? Form and social function in facial expressions. *Personality and Social Psychology Bulletin*, *31*(1), 73–86. <https://doi.org/10.1177/0146167204271306>
- McCullough, M. E., & Reed, L. I. (2016). What the face communicates: Clearing the conceptual ground. *Current Opinion in Psychology*, *7*, 110–114. <https://doi.org/10.1016/j.copsyc.2015.08.023>

- McKenzie, C. R. , & Liersch, M. J. (2011). Misunderstanding savings growth: Implications for retirement savings behavior. *Journal of Marketing Research*, *48*, S1–S13.  
<https://doi.org/10.1509/jmkr.48.SPL.S1>
- Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, *90*(5), 862–877. <https://doi.org/10.1037/0022-3514.90.5.862>
- Mehu, M., Little, A. C., & Dunbar, R. I. M. (2007). Duchenne smiles and the perception of generosity and sociability in faces. *Journal of Evolutionary Psychology*, *5*, 183–196.  
<https://doi.org/10.1556/JEP.2007.1011>
- Messias, J., Vikatos, P., & Benevenuto, F. (2017). White, Man, and highly followed: Gender and race inequalities in Twitter. *IEEE/WIC/ACM International Conference on Web Intelligence*.  
<https://doi.org/10.1145/3106426.3106472>
- Mieth, L., Bell, R., & Buchner, A. (2016). Cognitive load does not affect the behavioral and cognitive foundations of social cooperation. *Frontiers in Psychology*, *7*, 1–14.  
<https://doi.org/10.3389/fpsyg.2016.01312>
- Miller, A. H., Wattenberg, M. P., & Malanchuk, O. (1986). Schematic assessments of presidential candidates. *The American Political Science Review*, *80*(2), 521–540.
- Morewedge, C. K., & Kahneman, D. (2017). Associative processes in intuitive judgment. *Trends in Cognitive Sciences*, *14*(10), 435–440.  
<https://doi.org/10.1016/j.tics.2010.07.004>
- Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 129–140.  
<https://doi.org/10.1177/2372732215600886>
- Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for common designs*. R package version 0.9.12-4.1. Retrieved from <https://cran.r-project.org/package=BayesFactor>
- Mosca, L. (2014). The Five Star Movement: Exception or vanguard in Europe? *The International Spectator*, *49*(1), 36–52.  
<https://doi.org/10.1080/03932729.2013.875821>

- Murphy, L. (2016). *Airbnb's work to fight discrimination and build inclusion: A report submitted to Airbnb*. Retrieved from [http://blog.airbnb.com/wp-content/uploads/2016/09/REPORT\\_Airbnbs-Work-to-Fight-Discrimination-and-Build-Inclusion.pdf](http://blog.airbnb.com/wp-content/uploads/2016/09/REPORT_Airbnbs-Work-to-Fight-Discrimination-and-Build-Inclusion.pdf)
- Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin*, *35*(12), 1661–1671. <https://doi.org/10.1177/0146167209346309>
- Neuberg, S. L., & Fiske, S. T. (1987). Motivational influences on impression formation: Outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology*, *53*(3), 431–444. <https://doi.org/10.1037//0022-3514.53.3.431>
- Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *43*(7), 1154–1170. <https://doi.org/10.1037/xlm0000372>
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. <https://doi.org/10.1037/1089-2680.2.2.175>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- Oh, D., Buck, E. A., & Todorov, A. (2019). Revealing hidden gender biases in competence impressions of faces. *Psychological Science*, *30*(1), 65–79. <https://doi.org/10.1177/0956797618813092>
- Olivola, C. Y., Eubanks, D. L., & Lovelace, J. B. (2014). The many (distinctive) faces of leadership: Inferring leadership domain from facial appearance. *The Leadership Quarterly*, *25*(5), 817–834. <https://doi.org/10.1016/j.leaqua.2014.06.002>
- Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, *18*(11), 566–570. <https://doi.org/10.1016/j.tics.2014.09.007>

- Olivola, C. Y., Tingley, D., & Todorov, A. (2018). Republican voters prefer candidates who have conservative-looking faces: New evidence from exit polls. *Political Psychology, 39*(5), 1157–1171. <https://doi.org/10.1111/pops.12489>
- Olivola, C. Y., & Todorov, A. (2010a). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior, 34*(2), 83–110. <https://doi.org/10.1007/s10919-009-0082-1>
- Olivola, C. Y., & Todorov, A. (2010b). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology, 46*(2), 315–324. <https://doi.org/10.1016/j.jesp.2009.12.002>
- 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>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). <https://doi.org/10.1126/science.aac4716>
- Palermo, R., & Rhodes, G. (2007). Are you always on my mind? A review of how face perception and attention interact. *Neuropsychologia, 45*(1), 75–92. <https://doi.org/10.1016/j.neuropsychologia.2006.04.025>
- Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23*(3), 184–188. <https://doi.org/10.1177/0963721414531598>
- Parker, L. R., Monteith, M. J., Moss-Racusin, C. A., & Van Camp, A. R. (2018). Promoting concern about gender bias with evidence-based confrontation. *Journal of Experimental Social Psychology, 74*, 8–23. <https://doi.org/10.1016/j.jesp.2017.07.009>
- Payne, J., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*(3), 522–534.
- Pazda, A. D., Thorstenson, C. A., Elliot, A. J., & Perrett, D. I. (2016). Women's facial redness increases their perceived attractiveness: Mediation through perceived healthiness. *Perception, 45*(7), 739–754. <https://doi.org/10.1177/0301006616633386>



- Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2015). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods*, *48*(1), 341–348.  
<https://doi.org/10.3758/s13428-015-0576-1>
- Penton-Voak, I. S., Pound, N., Little, A. C., & Perrett, D. I. (2006). Personality judgments from natural and composite facial images: More evidence for a “kernel of truth” in social perception. *Social Cognition*, *24*(5), 607–640.  
<https://doi.org/10.1521/soco.2006.24.5.607>
- Perrett, D. I., Burt, D. M., Penton-Voak, I. S., Lee, K. J., Rowland, D. a, & Edwards, R. (1999). Symmetry and human facial attractiveness. *Evolution and Human Behavior*, *20*, 295–307.  
[https://doi.org/10.1016/S1090-5138\(99\)00014-8](https://doi.org/10.1016/S1090-5138(99)00014-8)
- Peters, J. G., & Welch, S. (1980). The effects of charges of corruption on voting behavior in congressional elections. *The American Political Science Review*, *74*(3), 697–708.
- Peters, M., Rhodes, G., & Simmons, L. W. (2007). Contributions of the face and body to overall attractiveness. *Animal Behaviour*, *73*(6), 937–942. <https://doi.org/10.1016/j.anbehav.2006.07.012>
- Petsko, C. D., & Bodenhausen, G. V. (2019). Race–crime congruency effects revisited: Do we take defendants’ sexual orientation into account? *Social Psychological and Personality Science*, *10*(1), 73–81. <https://doi.org/10.1177/1948550617736111>
- Pinkham, A. E., Griffin, M., Baron, R., Sasson, N. J., & Gur, R. C. (2010). The face in the crowd effect: Anger superiority when using real faces and multiple identities. *Emotion*, *10*(1), 141–146.  
<https://doi.org/10.1037/a0017387>
- Plaks, J. E., Stroessner, S. J., Dweck, C. S., & Sherman, J. W. (2001). Person theories and attention allocation: Preferences for stereotypic versus counterstereotypic information. *Journal of Personality and Social Psychology*, *80*(6), 876–893.  
<https://doi.org/10.1037//0022-3514.80.6.876>
- Porter, S., ten Brinke, L., & 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>

- Quattrone, G. A., & Tversky, A. (1988). Contrasting rational and psychological analyses of political choice. *American Political Science Review*, 82(3), 719–736.
- R Core Team. (2019). *R: A language and environment for statistical computing*. Retrieved from <https://www.r-project.org/>
- Radke, S., Kalt, T., Wagels, L., & Derntl, B. (2018). Implicit and explicit motivational tendencies to faces varying in trustworthiness and dominance in men. *Frontiers in Behavioral Neuroscience*, 12(8), 1–10. <https://doi.org/10.3389/fnbeh.2018.00008>
- Ravina, E. (2008). *Love & loans: The effect of beauty and personal characteristics in credit markets*. Retrieved from [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1101647](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1101647)
- Re, D. E., & Rule, N. (2017). Distinctive facial cues predict leadership rank and selection. *Personality and Social Psychology Bulletin*, 43(9), 1311–1322. <https://doi.org/10.1177/0146167217712989>
- Régner, I., Thinus-Blanc, C., Netter, A., Schmader, T., & Huguet, P. (2019). Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature Human Behaviour*. <https://doi.org/10.1038/s41562-019-0686-3>
- Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLoS ONE*, 7(3), e34293. <https://doi.org/10.1371/journal.pone.0034293>
- Rezlescu, C., Penton, T., Walsh, V., Tsujimura, H., Scott, S. K., & Banissy, M. J. (2015). Dominant voices and attractive faces: The contribution of visual and auditory information to integrated person impressions. *Journal of Nonverbal Behavior*, 39(4), 355–370. <https://doi.org/10.1007/s10919-015-0214-8>
- Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology*, 57, 199–226. <https://doi.org/10.1146/annurev.psych.57.102904.190208>
- Rhodes, G., Simmons, L. W., & Peters, M. (2005). Attractiveness and sexual behavior: Does attractiveness enhance mating success? *Evolution and Human Behavior*, 26, 186–201. <https://doi.org/10.1016/j.evolhumbehav.2004.08.014>

- Rhodes, G., Yoshikawa, S., Palermo, R., Simmons, L. W., Peters, M., Lee, K., ... Crawford, J. R. (2007). Perceived health contributes to the attractiveness of facial symmetry, averageness, and sexual dimorphism. *Perception, 36*, 1244–1253.  
<https://doi.org/10.1068/p5712>
- Rhue, L., & Clark, J. (2016). *Who gets started on Kickstarter? Racial disparities in crowdfunding success*. Retrieved from  
<http://www.ssrn.com/abstract=2837042>
- Riggio, H. R., & Riggio, R. E. (2002). Emotional expressiveness, extraversion, and neuroticism: A meta-analysis. *Journal of Nonverbal Behavior, 26*(4), 195–218.  
<https://doi.org/10.1023/A:1022117500440>
- Ritchie, K. L., Palermo, R., & Rhodes, G. (2017). Forming impressions of facial attractiveness is mandatory. *Scientific Reports, 7*(469), 1–8.  
<https://doi.org/10.1038/s41598-017-00526-9>
- Ro, T., Russell, C., & Lavie, N. (2001). Changing faces: A detection advantage in the flicker paradigm. *Psychological Science, 12*(1), 94–99. <https://doi.org/10.1111/1467-9280.00317>
- Roger, D., & Nesselroever, W. (1987). The construction and preliminary validation of a scale for measuring emotional control. *Personality and Individual Differences, 8*(4), 527–534.  
[https://doi.org/10.1016/0191-8869\(87\)90215-7](https://doi.org/10.1016/0191-8869(87)90215-7)
- Rosar, U., Klein, M., & Beckers, T. (2008). The frog pond beauty contest: Physical attractiveness and electoral success of the constituency candidates at the North Rhine-Westphalia state election of 2005. *European Journal of Political Research, 47*(1), 64–79.  
<https://doi.org/10.1111/j.1475-6765.2007.00720.x>
- Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34–55. <https://doi.org/10.1086/260169>
- Ruffle, B. J., & Shtudiner, Z. (2015). Are good-looking people more employable? *Management Science, 61*(8), 1760–1776.  
<https://doi.org/10.1287/mnsc.2014.1927>
- Rule, N. O., & Ambady, N. (2008). The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological Science, 19*(2), 109–111.  
<https://doi.org/10.1111/j.1467-9280.2008.02054.x>

- Rule, N. O., Ambady, N., Adams, R. B., Ozone, H., Nakashimi, S., Yoshikawa, S., & Watabe, M. (2010). Polling the face: Prediction and consensus across cultures. *Journal of Personality and Social Psychology, 98*(1), 1–15. <https://doi.org/10.1037/a0017673>
- Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology, 104*(3), 409–426. <https://doi.org/10.1037/a0031050>
- Rupp, D. E., Vodanovich, S. J., & Credé, M. (2006). Age bias in the workplace: The impact of ageism and causal attributions. *Journal of Applied Social Psychology, 36*(6), 1337–1364. <https://doi.org/10.1111/j.0021-9029.2006.00062.x>
- Russell, R., Porcheron, A., Sweda, J. R., Jones, A. L., Mauger, E., & Morizot, F. (2016). Facial contrast is a cue for perceiving health from the face. *Journal of Experimental Psychology: Human Perception and Performance, 42*(9), 1354–1362. <https://doi.org/10.1037/xhp0000219>
- Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion, 9*(2), 260–264. <https://doi.org/10.1037/a0014681>
- Satchell, L. P. (2019). From photograph to face-to-face: Brief interactions change person and personality judgments. *Journal of Experimental Social Psychology, 82*, 266–276. <https://doi.org/10.1016/j.jesp.2019.02.010>
- Satchell, L. P., Davis, J. P., Julle-Danière, E., Tupper, N., & Marshman, P. (2018). Recognising faces but not traits: Accurate personality judgment from faces is unrelated to superior face memory. *Journal of Research in Personality, 79*, 49–58.
- Schaller, M., & Duncan, L. A. (2007). The behavioral immune system. In J. P. Forgas, M. G. Haselton, & W. von Hippel (Eds.), *The handbook of evolutionary psychology* (pp. 293–307). <https://doi.org/10.1002/9781119125563.evpsych107>
- Scherpenzeel, A. C., & Das, M. (2010). “True” longitudinal and probability-based internet panels: Evidence from the netherlands. In M. Das, P. Ester, & K. L. (Eds.), *Social and behavioral research and the internet: Advances in applied methods and research strategies* (pp. 77–104). New York, NY: Taylor & Francis.

- Segal, N. L. (2013). Personality similarity in unrelated look-alike pairs: Addressing a twin study challenge. *Personality and Individual Differences, 54*(1), 23–28.  
<https://doi.org/10.1016/j.paid.2012.07.031>
- Segal, N. L., Graham, J. L., & Ettinger, U. (2013). Unrelated look-alikes: Replicated study of personality similarity and qualitative findings on social relatedness. *Personality and Individual Differences, 55*(2), 169–174. <https://doi.org/10.1016/j.paid.2013.02.024>
- Segal, N. L., Hernandez, B. A., Graham, J. L., & Ettinger, U. (2018). Pairs of genetically unrelated look-alikes: Further tests of personality similarity and social affiliation. *Human Nature, 29*, 402–417.  
<https://doi.org/10.1007/s12110-018-9326-2>
- Sellier, A.-L., Scopelliti, I., & Morewedge, C. K. (2019). Debiasing training transfers to improve decision making in the field. *Psychological Science, 30*(9), 1371–1379.  
<https://doi.org/10.1177/0956797619861429>
- Shah, A. K. (2007). Easy does it: The role of fluency in cue weighting. *Judgment and Decision Making, 2*(6), 371–379.  
<https://doi.org/10.1037/e722852011-015>
- Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin, 134*(2), 207–222. <https://doi.org/10.1037/0033-2909.134.2.207>
- Shariff, A. F., & Tracy, J. L. (2011). What are emotion expressions for? *Current Directions in Psychological Science, 20*(6), 395–399.  
<https://doi.org/10.1177/0963721411424739>
- Shevlin, M., Walker, S., Davies, M. N. O., Banyard, P., & Lewis, C. A. (2003). Can you judge a book by its cover? Evidence of self-stranger agreement on personality at zero acquaintance. *Personality and Individual Differences, 35*(6), 1373–1383.  
[https://doi.org/10.1016/S0191-8869\(02\)00356-2](https://doi.org/10.1016/S0191-8869(02)00356-2)
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.
- Simmons, J. P., & Nelson, L. D. (2006). Intuitive confidence: Choosing between intuitive and nonintuitive alternatives. *Journal of Experimental Psychology: General, 135*(3), 409–428.  
<https://doi.org/10.1037/0096-3445.135.3.409>

- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366.  
<https://doi.org/10.1177/0956797611417632>
- Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1), 99–118.  
<https://doi.org/10.2307/1884852>
- Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26(5), 559–569. <https://doi.org/10.1177/0956797614567341>
- Slepian, M. L., & Ames, D. R. (2015). Internalized impressions: The link between apparent facial trustworthiness and deceptive behavior is mediated by targets' expectations of how they will be judged. *Psychological Science*, 27(2), 282–288.  
<https://doi.org/10.1177/0956797615594897>
- Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgement. *Organizational Behavior and Human Performance*, 6, 649–744.
- Snijders, C. C. P. C., & Keren, G. G. (1999). Determinants of trust. In D. V Budescu, I. Erev, & R. Zwick (Eds.), *Games and human behavior: Essays in honor of Amnon Rapoport* (pp. 355–385). Mahwah, NJ: Erlbaum.
- Soll, J. B., Milkman, K. L., & Payne, J. W. (2014). A user's guide to debiasing. In G. Wu & G. Keren (Eds.), *Handbook of judgment and decision-making*. New York, NY: Wiley.
- Spisak, B. R., Dekker, P. H., Krüger, M., & Van Vugt, M. (2012). Warriors and peacekeepers: Testing a biosocial implicit leadership hypothesis of intergroup relations using masculine and feminine faces. *PLoS ONE*, 7(1), e30399.  
<https://doi.org/10.1371/journal.pone.0030399>
- Sprengelmeyer, R., Young, A. W., Baldas, E.-M., Ratheiser, I., Sutherland, C. A. M., Müller, H.-P., ... Orth, M. (2016). The neuropsychology of first impressions: Evidence from Huntington's disease. *Cortex*.  
<https://doi.org/10.1016/j.cortex.2016.10.006>
- Starr, S. B. (2014). Estimating gender disparities in federal criminal cases. *American Law and Economics Review*, 17(1), 127–159.  
<https://doi.org/10.1093/aler/ahu010>

- Steffensmeier, D., Ulmer, J., & Kramer, J. (1998). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. *Criminology*, *36*(4), 763–798. <https://doi.org/10.1111/j.1745-9125.1998.tb01265.x>
- Stewart, L. H., Ajina, S., Getov, S., Bahrami, B., Todorov, A., & Rees, G. (2012). Unconscious evaluation of faces on social dimensions. *Journal of Experimental Psychology: General*, *141*(4), 715–727. <https://doi.org/10.1037/a0027950>
- Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science*, *21*(3), 349–354. <https://doi.org/10.1177/0956797610362647>
- Stoker, J. I., Garretsen, H., & Spreeuwes, L. J. (2016). The facial appearance of CEOs: Faces signal selection but not performance. *PloS ONE*, *11*(7), e0159950. <https://doi.org/10.1371/journal.pone.0159950>
- Stolier, R. M., Hehman, E., & Freeman, J. B. (2018). A dynamic structure of social trait space. *Trends in Cognitive Sciences*, *22*(3), 197–200. <https://doi.org/10.1016/j.tics.2017.12.003>
- Stolier, R. M., Hehman, E., & Freeman, J. B. (2020). Trait knowledge forms a common structure across social cognition. *Nature Human Behaviour*. <https://doi.org/10.1038/s41562-019-0800-6>
- Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.1807222115>
- Sullivan, J. (2018). The primacy effect in impression formation: Some replications and extensions. *Social Psychological and Personality Science*, *10*(4), 1–8. <https://doi.org/10.1177/1948550618771003>
- Sussman, A. B., Petkova, K., & Todorov, A. (2013). Competence ratings in US predict presidential election outcomes in Bulgaria. *Journal of Experimental Social Psychology*, *49*(4), 771–775. <https://doi.org/10.1016/j.jesp.2013.02.003>
- Sutherland, C. A. M., Liu, X., Zhang, L., Chu, Y., Oldmeadow, J. A., & Young, A. (2017). Facial first impressions across culture: Data-driven modelling of Chinese and British perceivers' unconstrained facial impressions. *Personality and Social Psychology Bulletin*, *44*(4), 521–537. <https://doi.org/10.1177/0146167217744194>

- Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118.  
<https://doi.org/10.1016/j.cognition.2012.12.001>
- Sutherland, C. A. M., Oldmeadow, J. A., & Young, A. W. (2016). Integrating social and facial models of person perception: Converging and diverging dimensions. *Cognition*, *157*, 257–267.  
<https://doi.org/10.1016/j.cognition.2016.09.006>
- Sutherland, C. A. M., Rhodes, G., Burton, N., & Young, A. W. (2019). Do facial first impressions reflect a shared social reality? *British Journal of Psychology*. <https://doi.org/10.1111/bjop.12390>
- Sutherland, C. A. M., Young, A. W., & Rhodes, G. (2017). Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties. *British Journal of Psychology*, *108*(2), 397–415.  
<https://doi.org/10.1111/bjop.12206>
- Suzuki, A. (2016). Persistent reliance on facial appearance among older adults when judging someone's trustworthiness. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *73*(4), 573–583. <https://doi.org/10.1093/geronb/gbw034>
- Suzuki, A., Tsukamoto, S., & Takahashi, Y. (2017). Faces tell everything in a just and biologically determined world. *Social Psychological and Personality Science*, *10*(1), 62–72.  
<https://doi.org/10.1177/1948550617734616>
- Sylwester, K., Lyons, M., Buchanan, C., Nettle, D., & Roberts, G. (2012). The role of Theory of Mind in assessing cooperative intentions. *Personality and Individual Differences*, *52*(2), 113–117.  
<https://doi.org/10.1016/j.paid.2011.09.005>
- Tabak, J. A., & Zayas, V. (2012). The roles of featural and configural face processing in snap judgments of sexual orientation. *PLoS ONE*, *7*(5), e36671. <https://doi.org/10.1371/journal.pone.0036671>
- Tamir, D. I., & Mitchell, J. P. (2012). Anchoring and adjustment during social inferences. *Journal of Experimental Psychology: General*, *142*(1), 151–162. <https://doi.org/10.1037/a0028232>
- Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven & London: Yale University Press.



- Theeuwes, J., & Van der Stigchel, S. (2006). Faces capture attention: Evidence from inhibition of return. *Visual Cognition*, *13*(6), 657–665. <https://doi.org/10.1080/135062805000410949>
- Thielmann, I., & Hilbig, B. E. (2014). Trust in me, trust in you: A social projection account of the link between personality, cooperativeness, and trustworthiness expectations. *Journal of Research in Personality*, *50*(1), 61–65. <https://doi.org/10.1016/j.jrp.2014.03.006>
- Thielmann, I., & Hilbig, B. E. (2015). Trust: An integrative review from a person-situation perspective. *Review of General Psychology*, *19*(3), 249–277. <https://doi.org/10.1037/gpr0000046>
- Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113. <https://doi.org/10.1037/t49856-000>
- Thorstenon, C. A., & Elliot, A. J. (2017). Subjective perception of color differences is greater for faces than non-faces. *Social Cognition*, *35*(3), 299–312. <https://doi.org/10.1521/soco.2017.35.3.299>
- Thorstenon, C. A., Pazda, A. D., Elliot, A. J., & Perrett, D. I. (2017). Facial redness increases men's perceived healthiness and attractiveness. *Perception*, *46*(6), 650–664. <https://doi.org/10.1177/0301006616680124>
- Tiddeman, B., Burt, M., & Perrett, D. I. (2001). Prototyping and transforming facial textures for perception research. *Computer Graphics and Applications, IEEE*, *21*(5), 42–50.
- Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton: Princeton University Press.
- Todorov, A., Funk, F., & Olivola, C. Y. (2015). Response to Bonnefon et al.: Limited 'kernels of truth' in facial inferences. *Trends in Cognitive Sciences*, *19*(8), 422. <https://doi.org/10.1016/j.tics.2015.05.002>
- Todorov, A., Loehr, V., & Oosterhof, N. N. (2010). The obligatory nature of holistic processing of faces in social judgments. *Perception*, *39*(4), 514–532. <https://doi.org/10.1068/p6501>
- Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623–1626. <https://doi.org/10.1126/science.1110589>

- Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology, 66*(1), 519–545. <https://doi.org/10.1146/annurev-psych-113011-143831>
- Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27*(6), 813–833. <https://doi.org/10.1521/soco.2009.27.6.813>
- Todorov, A., & Porter, J. M. (2014a). Misleading first impressions: Different for different facial images of the same person. *Psychological Science, 25*(7), 1404–1417. <https://doi.org/10.1177/0956797614532474>
- Todorov, A., & Porter, J. M. (2014b). Misleading first impressions: Different for different facial images of the same person Supplemental Material. *Psychological Science, 25*(7), 1404–1417. <https://doi.org/10.1177/0956797614532474>
- Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12*(12), 455–460. <https://doi.org/10.1016/j.tics.2008.10.001>
- Tognetti, A., Berticat, C., Raymond, M., & Faurie, C. (2013). Is cooperativeness readable in static facial features? An inter-cultural approach. *Evolution and Human Behavior, 34*(6), 427–432. <https://doi.org/10.1016/j.evolhumbehav.2013.08.002>
- Tsankova, E., Krumhuber, E., Aubrey, A. J., Kappas, A., Möllering, G., & Rosin, P. L. (2015). The multi-modal nature of trustworthiness perception. *Proceedings of the International Speech Communication Association (ISCA)*, 147–152.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.
- van't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition, 108*, 796–803. <https://doi.org/10.1016/j.cognition.2008.07.002>
- Van De Ven, N., Bogaert, A., Serlie, A., Brandt, M. J., Denissen, J. J. A., & Serlie, A. (2017). Personality perception based on LinkedIn profiles. *Journal of Managerial Psychology, 32*(6), 418–429. <https://doi.org/10.1108/JMP-07-2016-0220>

- Van Kleef, G. A. (2010). The emerging view of emotion as social information. *Social and Personality Psychology Compass*, 4(5), 331–343. <https://doi.org/10.1111/j.1751-9004.2010.00262.x>
- Verhulst, B., Lodge, M., & Lavine, H. (2010). The attractiveness halo: Why some candidates are perceived more favorably than others. *Journal of Nonverbal Behavior*, 34(2), 111–117. <https://doi.org/10.1007/s10919-009-0084-z>
- Verplaetse, J., Vanneste, S., & Braeckman, J. (2007). You can judge a book by its cover: The sequel. A kernel of truth in predictive cheating detection. *Evolution and Human Behavior*, 28(4), 260–271. <https://doi.org/10.1016/j.evolhumbehav.2007.04.006>
- Vogt, S., Efferson, C., & Fehr, E. (2013). Can we see inside? Predicting strategic behavior given limited information. *Evolution and Human Behavior*, 34(4), 258–264. <https://doi.org/10.1016/j.evolhumbehav.2013.03.003>
- Waits, C., & Little, A. C. (2006). Preferences for symmetry in conspecific facial shape among *Macaca mulatta*. *International Journal of Primatology*, 27(1), 133–145. <https://doi.org/10.1007/s10764-005-9015-y>
- Waits, C., Little, A. C., Wolfensohn, S., Honess, P., Brown, A. P., Buchanan-Smith, H. M., & Perrett, D. I. (2003). Evidence from rhesus macaques suggests that male coloration plays a role in female primate mate choice. *Proceedings of the Royal Society B: Biological Sciences*, 270, S144–S146. <https://doi.org/10.1098/rsbl.2003.0065>
- Walker, M., Wänke, M., Davies, M., Banyard, P., Lewis, C., & Abdi, H. (2017). Caring or daring? Exploring the impact of facial masculinity/femininity and gender category information on first impressions. *PloS ONE*, 12(10), e0181306. <https://doi.org/10.1371/journal.pone.0181306>
- Wang, D., Xi, S., & Gilheany, J. (2015). The model minority? Not on Airbnb.com: A hedonic pricing model to quantify racial bias against Asian Americans. *Technology Science*. Retrieved from <http://techscience.org/a/2015090104/#Suggestions>
- Wedel, M., & Pieters, R. (2007). A review of eye-tracking research in marketing. *Review of Marketing Research*, 4, 123–147. [https://doi.org/10.1108/S1548-6435\(2008\)0000004009](https://doi.org/10.1108/S1548-6435(2008)0000004009)

- Welch, S., & Hibbing, J. R. (1997). The effects of charges of corruption on voting behavior in congressional elections. *The Journal of Politics*, *59*(1), 226–239.
- Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598. <https://doi.org/10.1111/j.1467-9280.2006.01750.x>
- 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, R. K., & Eckel, C. C. (2006). Judging a book by its cover: Beauty and expectations in the trust game. *Political Research Quarterly*, *59*(2), 189–202. <https://doi.org/10.1177/106591290605900202>
- Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*(1), 117–142. <https://doi.org/10.1037/0033-2909.116.1.117>
- Winston, J. S., Strange, B. A., O’Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience*, *5*(3), 277–283. <https://doi.org/10.1038/nn816>
- Woods, K. (2017). “Facing” identity in a “faceless” society: Physiognomy, facial appearance and identity perception in eighteenth-century London. *Cultural and Social History*, *14*(2), 137–153. <https://doi.org/10.1080/14780038.2017.1290998>
- Wu, X., & Zhang, X. (2016). *Automated inference on criminality using face images*. Retrieved from <http://arxiv.org/abs/1611.04135>
- Xie, S. Y., Flake, J. K., & Hehman, E. (2019). Perceiver and target characteristics contribute to impression formation differently across race and gender. *Journal of Personality and Social Psychology*, *117*(2), 364–385. <https://doi.org/10.1037/pspi0000160>
- Yu, M., Saleem, M., & Gonzalez, C. (2014). Developing trust: First impressions and experience. *Journal of Economic Psychology*, *43*, 16–29. <https://doi.org/10.1016/j.joep.2014.04.004>

- Yukl, G. (1989). Managerial leadership: A review of theory and research. *Journal of Management*, 15(2), 251–289.
- Zakay, D. (1985). Post-decisional confidence and conflict experienced in a choice process. *Acta Psychologica*, 58, 75–80.
- Zebrowitz, L. A. (2012). Ecological and social approaches to face perception. In G. Rhodes, A. Calder, M. Johnson, & J. V. Haxby (Eds.), *Oxford handbook of face perception*.  
<https://doi.org/10.1093/oxfordhb/9780199559053.013.0003>
- Zebrowitz, L. A. (2017). First impressions from faces. *Current Directions in Psychological Science*, 26(3), 237–242.  
<https://doi.org/10.1177/0963721416683996>
- Zebrowitz, L. A., Fellous, J. M., Mignault, A., & Andreoletti, C. (2003). Trait impressions as overgeneralized responses to adaptively significant facial qualities: Evidence from connectionist modeling. *Personality and Social Psychology Review*, 7(3), 194–215.  
[https://doi.org/10.1207/S15327957PSPR0703\\_01](https://doi.org/10.1207/S15327957PSPR0703_01)
- 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. M. (1992). Impressions of babyfaced males and females across the lifespan. *Developmental Psychology*, 28(6), 1143–1152.
- Zebrowitz, L. A., & Rhodes, G. (2004). Sensitivity to “bad genes” and the anomalous face overgeneralization effect: Cue validity, cue utilization, and accuracy in judging intelligence and health. *Journal of Nonverbal Behavior*, 28(3), 167–185.  
<https://doi.org/10.1023/B:JONB.0000039648.30935.1b>



# Acknowledgments

There are many people without whom this dissertation would have never been written.

My supervisors, Ilja, Tony, and Mariëlle. You showed me the ropes and taught me a lot about what it means to be a good researcher. You gave me the freedom to pursue my own interests (which I appreciated a lot), but also the reassuring feeling that I could always come to you with any kind of question. Thanks for all the guidance in the last four years and I'm looking forward to continue working with you in the future.

I would also like to thank everyone else in the Social Psychology department. Be it the weird lunch conversations, the feedback given during lab meetings, the crazy ASPO parties and department outings, or the general welcoming atmosphere that made me feel like I could walk into anybody's office to strike up a conversation about anything at any time: All of you contributed to a work environment that was extremely fun to work in.

A special thanks goes to all the other PhD students that were beside me in the metaphorical trenches over the last years. Thanks for creating a collaborative and supportive environment that featured not only stimulating discussions about all sorts of topics, but also Nerf guns, ping pong battles, fun nights with movies and games, and of course hundreds of "coooffeEEEE" breaks.

During my time as a PhD student, I had the privilege to collaborate with a number of people outside the department, including Alex Todorov, Philipp Gerlach, Alex Jones, and Liam Satchell. You all broadened my academic horizon, showed me cool new methods, and I'm looking forward to working with you in the future. A special thanks goes to Alex Todorov who welcomed me in Princeton for a very productive and fun research visit. Some other people also deserve a mention here, either because they served as role models for what kind of researcher I would like to become, or because their research continues to amaze and inspire me (or both). These people include, among many others: Lisa



DeBruine, Ben Jones, Ron Dotsch, Clare Sutherland, Jon Freeman, Eric Hehman, Nick Rule, Jean-François Bonnefon, Chris Olivola, and Leslie Zebrowitz. Even though I did not have the chance to collaborate or chat with all of you yet, I hope that this will change in the future.

My three paranymphs: Fieke, Will, and Joeri. It's reassuring to have you stand behind me today because I know that I can always count on you.

Joeri, we shared an office (and a bottle of whisky) for the last four years. Friday afternoons are just not complete without a mandatory performance of "Don't Stop Believin'". I'm looking forward to more afternoons of work and (mostly) whisky, more game nights, and more discussions about beer!

Will, it was great to find another person who thinks doing research is a perfectly acceptable hobby and who does his best work after 6 pm. I hope that we continue our weekly bro nights even though I actually have to put on shoes to go to your place now. To more movie (and Skyjo) marathons, jeu de boules (and sock), ice cream runs, R courses, random research projects, and workations!

Fieke, my time in Tilburg not only gave me a PhD but also a partner in crime (and science and traveling and much more). I'm extremely lucky that I found someone as awesome as you. Someone who shares my interests in travel, food, science, and (of course) doggies. Someone who always supports me and makes me happy every day. Thanks for always being there, love, and I can't wait to further explore the world and life with you!

Also, thank you Fieke for introducing me to Bono (good girl!), and thanks to Toon and Myriam for always giving me unlimited access to her. En enorm bedankt dat jullie mij met open armen in jullie gezin verwelkomd hebben.

There are many people who made the last years in Tilburg a lot of fun. Thank you, Byron, Chrissy, Erdem, Flo, Gaby, Leonie, Lis, Maaïke,

Mich, Michael, Nina, Paulette, Rabia, Tine, Tünde and everyone else for welcoming me in Tilburg with open arms (despite my German-ness), for all the game and movie nights, Halloween and birthday parties, for helping me move (not once but twice), and much more.

Dank geht natürlich auch an meine Eltern und den Rest meiner Familie. Ihr habt mich über die Jahre hinweg immer in allen Vorhaben unterstützt. Auch wenn ihr das meiste, was in diesem Buch steht nicht ganz verstehen könnt, habt ihr einen sehr großen Teil dazu beigetragen, dass ich es schreiben kann. Ohne euren kontinuierlichen Beistand in den letzten, fast 30 Jahren hätte ich nicht zu diesem Punkt in meinem Leben kommen können.

**Kurt Lewin**  
**Institute**  
**dissertation series**

The “Kurt Lewin Institute dissertation series” started in 1997. Since 2016, the following dissertations have been published in this series:

2016-01: Anna van ‘t Veer: *Effortless morality — cognitive and affective processes in deception and its detection*

2016-02: Thijs Bouman: *Threat by association: How distant events can affect local intergroup relations*

2016-03: Tim Theeboom: *Workplace coaching: Processes and effects*

2016-04: Sabine Strofer: *Deceptive intent: Physiological reactions in different interpersonal contexts*

2016-05: Caspar van Lissa: *Exercising Empathy: The Role of Adolescents' Developing Empathy in Conflicts with Parents*

2016-06: Marlon Mooijman: *On the determinants and consequences of punishment goals: The role of power, distrust, and rule compliance*

2016-07: Niels van Doesum: *Social mindfulness*

2016-08: Leonie Venhoeven: *A look on the bright side of an environmentally-friendly life: Whether and why acting environmentally-friendly can contribute to well-being*

2016-09: Florien Cramwinckel: *The social dynamics of morality*

2016-10: Junhui Wu: *Understanding Human Cooperation: The Psychology of Gossip, Reputation, and Life History*

2016-11: Elise C. Seip: *Desire for vengeance. An emotion-based approach to revenge*

2016-12: Welmer E. Molenmaker: *The (un)willingness to reward cooperation and punish non-cooperation*

2016-13: Liesbeth Mann: *On Feeling Humiliated. The Experience of Humiliation in Interpersonal, Intragroup, and Intergroup Contexts*

2016-14: Angela M. Ruepert: *Working on the environment*

2016-15: Femke Hilverda: *Making sense of food risk information: The case of organic food.*

- 2016-16: Debora E. Purba: *Antecedents of turnover, organizational citizenship behavior, and workplace deviance: Empirical evidence from Indonesia.*
- 2016-17: Maja Kutlaca: *The Role of Values and Value-Identity Fit in Motivating Collective Action*
- 2016-18: Felicity Turner: *A New Psychological Perspective on Identity Content, its Conceptualization, Measurement, and Application*
- 2016-19: Tim W. Faber: *When Imitation Falls Short: The Case of Complementary Actions.*
- 2016-20: Daniela Becker: *Self-control conflict in the eating domain: A cognitive, affective and behavioral perspective*
- 2016-21: Zoi Manesi: *Prosocial Behavior Under Surveillance: Understanding the Eye-Images Effect*
- 2017-01: Tracy Cheung: *Turning vice into virtue - when low self-control states facilitate goal-oriented behaviours*
- 2017-02: Pum Kommattam: *Feeling the Other: Emotion Interpretation in Intercultural Settings*
- 2017-03: Lotte Veenstra: *Taming Tempers: A situated motivational approach to anger management*
- 2017-04: Jolien van Breen: *The path of most resistance: How groups cope with implicit social identity threat*
- 2017-05: Yuije Cheng: *Creativity Under the Gun: How Threat Features and Personal Characteristics Motivate Creative Responding*
- 2017-06: Eftychia Stamkou: *The dynamic nature of social hierarchies: The role of norm violations and hierarchical concerns*
- 2017-07: Anne Marthe van der Bles: *Societal Discontent - Deciphering the Zeitgeist*
- 2017-08: Willem Slegers: *Meaning and Pupillometry: The Role of Physiological Arousal in Meaning Maintenance*
- 2017-09: Julia Sasse: *More Than a Feeling: Strategic Emotion Expression in Intergroup Conflicts*
- 2017-10: Nils Köbis: *The Social Psychology of Corruption*

- 2017-11: Tim de Wilde: *Struggling to decide. Competition in group decision-making*
- 2017-12: Nathalie Boot: *The creative brain: Some insights into the neural dynamics of flexible and persistent creative processes*
- 2017-13: Johannes Seehusen: *Foregone and Forethought: Motivation in the Context of Past and Future Alternatives*
- 2017-14: Ernst Willem Meerholz: *The 'other' side of compassion: How the self avoids responsibility for past wrongs*
- 2017-15: Wieke Scholten: *Banking on Team Ethics: A team climate perspective on root causes of misconduct in financial services*
- 2018-01: Mike Keesman: *Observing the mind instead of acting on it: How mindfulness empowers people to live healthily*
- 2018-02: Marije Bakker: *Turning Crisis into Opportunity: the Influence of the Government and the Social environment*
- 2018-03: Miriam Oostinga: *Breaking (the) ice: Communication error management in law enforcement interactions*
- 2018-04: Xia Fang: *Perceiving and Producing Facial Expressions of Emotion: The Role of Dynamic Expressions and Culture*
- 2018-05: David Maij: *Sensing Supernatural Agency - An empirical quest on the socio-cognitive foundations of supernatural beliefs*
- 2018-06: Mariko Visserman: *The Art of Sacrifice: Self-Other Dilemmas, Biased Perceptions, and the Emergence of Gratitude*
- 2018-07: Caroline Schlinkert: *Minding the body: The role of rumination and stress in embodied information processing*
- 2018-08: Aafke van Mourik Broekman: *An Experimental Approach to Group Growth: When Boundaries Between Performers and Observers Are Breached*
- 2018-09: Judith Rachl: *Unconscious Bonding - Forming Bonds Quickly in Today's Fast-Paced Society*
- 2018-10: Bibiana Armenta Gutierrez: *Stepping into old age: A dynamic perspective on age identity change in the transition from midlife to older adulthood*

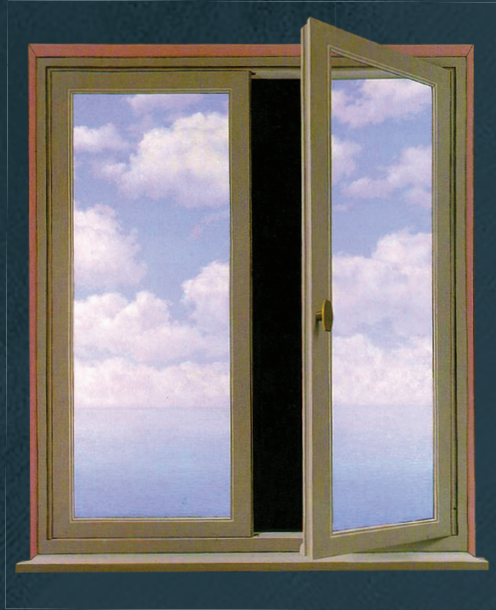
- 2018-11: Dalya Samur: *From reading to feeling: A language-based approach to alexithymia*
- 2018-12: Marloes Huis: *Women's empowerment in the context of microfinance services*
- 2018-13: Ernst Noppers: *Driving adoption: The symbolic value of sustainable innovations*
- 2018-14: Sosja Prinsen: *Justified indulgence: The effects of self-licensing on self-regulation over time*
- 2018-15: Ali Mashuri: *Dealing with Separatism Conflict in Indonesia: Examining an Interactive Model of Conflict De-Escalation and Resolution*
- 2018-16: Darya Moghimi: *Doing Well and Feeling Well: The role of Selection, Optimization, and Compensation as Strategies of Successful (Daily) Life Management*
- 2019-01: Wendy Schreurs: *Crossing Lines Together: How and why citizens participate in the police domain*
- 2019-02: Kiki de Jonge: *Stimulating Creativity: Matching Person and Context*
- 2019-03: Catherine Molho: *The Psychological Underpinnings of Cooperation and the Punishment of Non-Cooperators: Insights from the Lab to the Field*
- 2019-04: Xiaoyue Tan: *The Psychology of Loss Management*
- 2019-05: Lisanne Pauw: *A problem shared is a problem halved? On the dyadic nature of emotion regulation*
- 2019-06: Tina Venema: *Preferences as boundary condition of nudge effectiveness. The potential of nudges under empirical investigation*
- 2019-07: Loes Kreemers: *Searching for a Job: Problem- and Emotion-Focused Coping*
- 2019-08: Bastian Jaeger: *Facial Discrimination: The irresistible influence of first impressions*











kurtle

winins

tituut