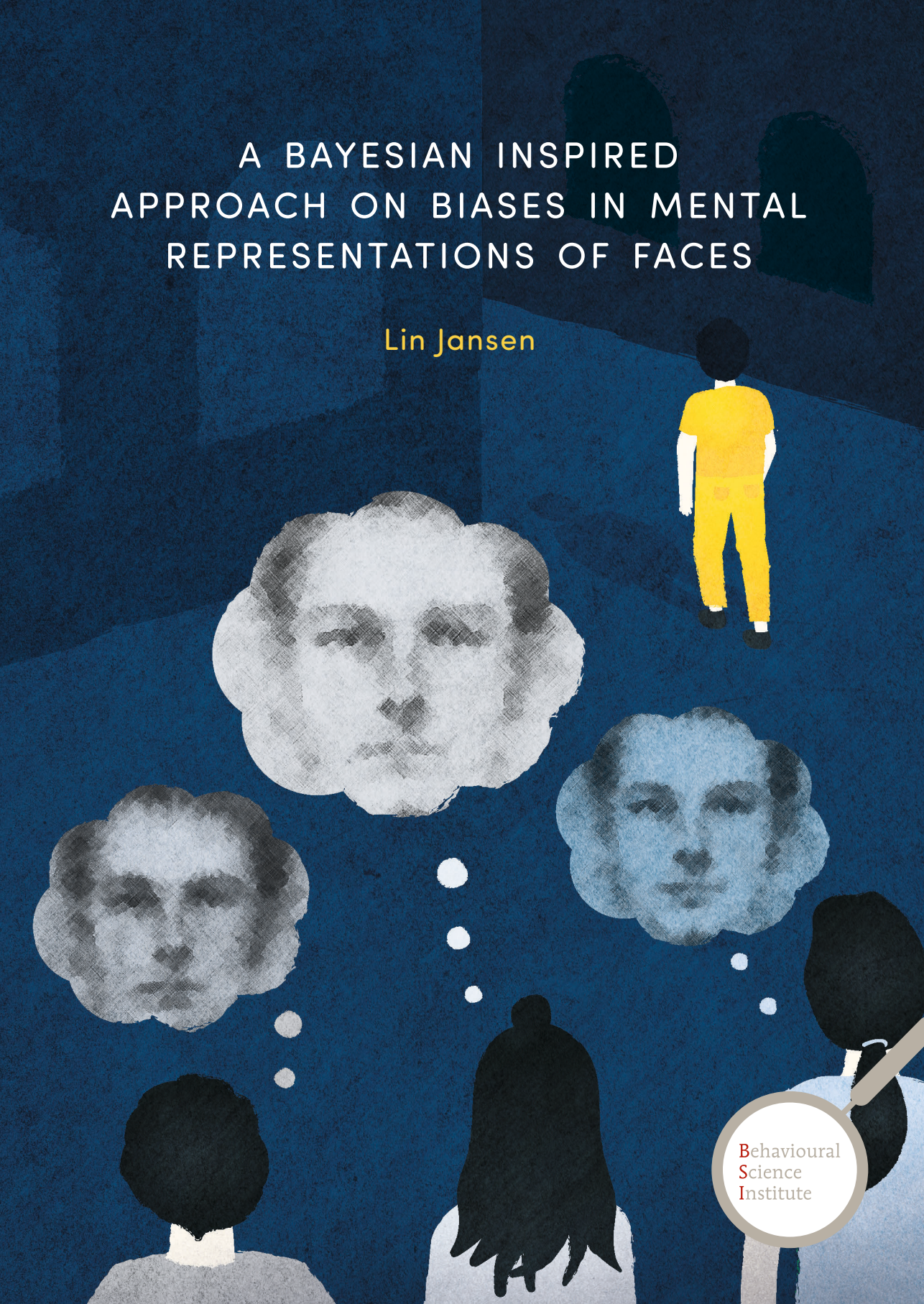
A BAYESIAN INSPIRED APPROACH ON BIASES IN MENTAL REPRESENTATIONS OF FACES

Lin Jansen



Behavioural Science Institute

A Bayesian Inspired Approach on Biases in Mental Representations of Faces

Lin Jansen

ISBN: 978-94-6458-107-2

Provided by thesis specialist Ridderprint, ridderprint.nl
Printing: Ridderprint
Layout and cover design: Sara Terwisscha van Scheltinga, persoonlijkproefschrift.nl

The research in this dissertation was supported by NWO grant 406-13-023 awarded to Lin Jansen MSc and prof. dr. Daniël Wigboldus.

© Lin Jansen, 2022. All rights reserved.

A Bayesian Inspired Approach on Biases in Mental Representations of Faces

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 12 april 2022
om 10.30 uur precies

door

Lin Fiene Jansen
geboren op 4 december 1988
te Arnhem

Promotoren:

prof. dr. D.H.J. Wigboldus

prof. dr. R.W. Holland

Copromotor:

dr. R. Dotsch (Anchormen)

Manuscriptcommissie:

prof. dr. I.J.E.I. van Rooij (voorzitter)

prof. dr. E.H. Gordijn (Rijksuniversiteit Groningen)

prof. dr. M. Rinck

Take small steps every day
Let your goals lead the way
Cultivate a sense of play
Keep friends that make you feel okay
And no matter how long the way
You will get there one day

CONTENTS


1

General Introduction

> "*Certainly for the most part, the way we see things is a combination of what is there and of what we expected to find.*" – Walter Lippmann (1922).

Imagine you want to go grocery shopping. When you approach the shopping center on foot, a group of people warns you not to go any further, because someone is robbing one of the stores. Chances are you start to visualize in your mind, whether consciously or unconsciously, what might be happening and what the perpetrator may look like. Indeed, research suggests that behavioral information creates a visual expectation of facial appearance (Dotsch, Wigboldus, & Van Knippenberg, 2013). Next, as the perpetrator flees the crime scene, you catch a glance of the perpetrator's face (which later leads to you being interviewed by the police as an eyewitness). Now that you have seen the actual face, what happens to your mental image of the face? Does your cognitive system overwrite the expectation from before by the visual input from the actual face, resulting in an accurate image of the perpetrator's face in your mind? Or is your mental representation of the perpetrator's face biased by the expectation you had beforehand, making it appear for example more criminal or untrustworthy looking than it actually was? And if so, which circumstances would decrease and increase the chance on such a biased mental representation?

Faces play an important role in people's lives. People attempt to extract all kinds of social information from other's facial appearance. They use faces to identify others as specific individuals, but also to infer information about a person's age, sex, race, emotions, intentions, and even personality (Sutherland et al., 2013; Todorov, 2017; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015; Zebrowitz, 2017). As such, impressions from faces lead to influential evaluations about the face bearer, for example whether the person is considered to be trustworthy or competent (Todorov, Dotsch, Porter, Oosterhof, & Falvello, 2013). The idea that the face can be used as a window to the soul has been documented already in the time of Aristotle and is still very much alive today (Todorov, 2017). An idea with dangerous implications, as history has shown. After the pseudoscience of physiognomy widely popularized the idea that a person's character can be read from the face, the idea fueled alarming ideologies, such as those propagated by Galton's eugenics societies and Nazi Germany (Todorov, 2017). Although physiognomy received no scientific support and has therefore long lost scientific credibility, people still form impressions from faces and

act upon these impressions everyday (Todorov, 2017). As such, physiognomic tendencies still contribute to problems of inequality in societies, reflected in exposed biases in the development, performance, and use of contemporary face recognition programs (Bacchini & Lorusso, 2019). How people perceive your face can thus have important social consequences, for example whether others trust you financially (van 't Wout & Sanfey, 2008) or to be the right fit for the job (Antonakis & Eubanks, 2017).

Yet, although faces attract much of people's attention (Fletcher-Watson, Findlay, Leekam, & Benson, 2008), people often already have information about a person before seeing his or her face for the first time. Just like in the crime scene example from above, but also in more prevalent real-life situations, for example through gossip, information online, application procedures, online communication, and legal procedures. Would such verbal information about a person shape people's mental representation of that person's face? Indeed, there is research suggesting that this is the case (e.g. Dotsch, Wigboldus, & Van Knippenberg, 2013). In fact, there is theoretical support for the idea that humans mentally represent knowledge perceptually, even abstract conceptual knowledge (Barsalou, 1999). Yet even then, would such verbal information still shape the visual mental representation of that person's face after the perceiver has actually seen the person's face? More specifically, under which circumstances and to what extent would such a bias (not) occur? These are exactly the kind of questions the current dissertation is about. Before diving into these questions, however, allow me to zoom out a bit and provide you with some relevant background information first. After all, why would our brains visually represent *anything* in a biased way? Would it not be far more useful if our visual mental representations were exact, flawless copies of the things that are really 'out there'?

FUNCTIONING IN A COMPLEX WORLD

Although having exact copies of the world 'out there' in our minds may sound ideal, it is flat out impossible. The world presents us with a wealth of information, which is too vast for our senses and brains to process in its entirety (Lippmann, 1922; Summerfield & Egner, 2009). To function in this complex world, our cognitive systems have to prioritize what to process (Allport,

1954; Barsalou, 1999). And even then, they usually cannot fully process every detail of the prioritized object(s). Moreover, the incoming information is often ambiguous: sensory input is noisy and can often be interpreted in multiple ways (Hohwy, Roepstorff, & Friston, 2008; Kersten & Yuille, 2003; Summerfield & Egner, 2009). For instance, an object as simple as a book produces different sensory input when viewed from different angles and under different lighting conditions. Additionally, a book may share visual properties with other objects that are not books (e.g. a tile or brick, a laptop or tablet, a picture frame, a cutting board, a briefcase), and yet differ in appearance from other books (e.g. in size, color, or material).

To function efficiently, our cognitive systems have to model a simplified version of the abundant and complex world (Allport, 1954; Barsalou, 1999; Bruner, 1957; Clark, 2013; Lippmann, 1922; Sherman, Lee, Bessenoff, & Frost, 1998). They summarize the multitude of objects and events into categories (like the category 'books') and adopt simplified rules about them and their relations (e.g. books are meant to be read, can often be found in libraries or on bookshelves in houses, are usually rectangular and made of paper, are small enough to hold with two hands, are usually someone's possession). Our cognitive systems spend our lives building and adapting such models, based on our own experiences and inferred beliefs and on those of others we have come to learn about.

These mental models are used to predict and interpret currently relevant – and often inherently ambiguous – information in the world (Barsalou, 1999; Clark, 2013; Edwards, Adams, Brown, Pareés, & Friston, 2012; Summerfield & Egner, 2009). For instance, imagine seeing something that looks like a book in the kitchen. Before having fully processed the object, you may predict with the help of your mental models that it is probably a cookbook and intend to open it, opposed to predicting it is probably a cutting board and intend to chop the onions on it. Our brains thus do not operate like passive recorders. They actively construct our reality with the help of the models in our minds and the contextual cues that are present in that moment.

Using these models to predict and interpret the world has advantages and disadvantages. As indicated above, actively predicting our environment prevents us from having to process every single detail, which saves a lot of time and

energy. Furthermore, we are quickly prepared for what might happen and how we should respond to that. If we predict well, we can respond fast and adequately and our interaction with the environment will run smoothly. Additionally, it allows us to more quickly identify unpredicted aspects in the environment, which is vital for our safety. However, our models are not always well suited to the situation, leading us to act upon inaccurate beliefs and perceptions. We sometimes consciously experience this, for instance when we are confronted with the consequences of misidentifying an object. As a Dutch teenager, I had been invited to eat dinner with friends from varying cultures. I confidently took a big bite out of what I assumed was a sweet pepper (the only kind of pepper I had ever eaten), only to find out that it was in fact a hot chili pepper. My eyes instantly watered up, my face turned red and sweaty, and my mouth felt like it was on fire. Needless to say, my model on pepper-like-looking foods changed drastically that day. Even within our familiar environments, though, our models may generate inaccurate beliefs leading to inaccurate perceptions (e.g. "I perceive my skin as looking healthier when I use the more expensive facial cream"). Yet, these can persist as long as we are not forced to adjust them. In sum, although definitely not perfect, the models used to inform our perception do save tremendous effort and allow us to survive in our world (Allport, 1954; Bruner, 1957; Clark, 2013; Lippmann, 1922).

Interestingly, appropriately interpreting *social* stimuli is even harder compared to non-social objects, because social stimuli are even more complex, fleeting, and ambiguous in nature (Bruner, 1957). It is impossible to directly observe another person's thoughts, feelings, or intentions. They have to be inferred. Consequently, it may often go unnoticed when interpretation is faulty because either there is no opportunity to further test the accuracy of the interpretation or the ambiguity of the social stimulus allows for an interpretation in line with the original expectation. For example, through continued perception you probably find out soon enough if the assumed cookbook in the example above is in fact a cutting board or vice versa. In comparison though, a crying athlete who just won second place is clearly perceived as crying but could still be interpreted as either sad or happy (Hassin, Aviezer, & Bentin, 2013; Medvec, Madey, & Gilovich, 1995). Or imagine that I notice someone walking behind me in a dark alley. I may predict that this person is probably dangerous and quickly flee the scene, never learning that this person was in fact a harmless passerby.

Or, I do not flee the scene, but interpret this person's behavior in line with my prediction, such that a simple nod is interpreted as an aggressive opposed to friendly signal (Galperin & Haselton, 2013).

Social stimuli are not only harder to correctly interpret compared to non-social stimuli, their interpretation can also have more far-reaching consequences. For instance, non-Black perceivers overestimate young Black (vs. White) men's physical size and strength, and are therefore likelier to justify the use of force against them when they behave aggressively (Wilson, Hugenberg, & Rule, 2017). Relatedly, police officers are likelier to misperceive unarmed Black (vs. White) suspects as armed and to shoot them (although this bias may potentially be eliminated through training; Plant & Peruche, 2005). Expectations based on biased beliefs about other people can thus lead to misperceptions with severe and sometimes tragic consequences.

As the opening quote of this chapter shows, the idea that our expectations actively influence our perception is not new. We also find this idea formalized in Bayesian models of perception (e.g. Mamassian, Landy, & Maloney, 2002). These models use conditional probabilities to show how both sensory input and expectations constrain perception (Zaki, 2013). For instance, consider that I am sitting in my bedroom and see a black dot on the wall. The probability that the black dot is a spider or P(spider | black dot) depends not only on the probability that I would be seeing a black dot if there is a spider or P(black dot | spider), but also on the prior probability that there is a spider on the bedroom wall or P(spider). If I believe I am very likely to find a spider in the bedroom (for example because my fear of spiders heightens my expectation to find one or because I saw a spider in the bedroom yesterday), I am likely to perceive the black dot as a spider, whereas my fiancé who does not expect to find a spider in the bedroom may interpret the black dot as the shadow of a dent in the wall.

Recently, the same idea has been applied in predictive coding models in neuroscience to describe on a neural level how the brain generates perception (Clark, 2013; Friston, 2010; Hohwy et al., 2008). These models state that each hierarchically higher neural level predicts activity in the neural level directly below it, based on the currently active prediction of what is expected to be 'out there'. At the same time, each lower neural level reports errors in prediction to

the level directly above it, adjusting the prediction at the higher level. These recurrent interactions continue on all levels of the hierarchy with the aim to minimize prediction error.

In sum, scientists from different approaches and centuries endorse the notion that human perception is an active reconstruction of what is 'out there', informed not only by sensory input from the world 'out there', but for a large part by the perceiver's expectations as well.

FACES AS IMPORTANT SOCIAL OBJECTS

As humans are social beings, this strategy of prioritizing, predicting and interpreting gets applied also to social interactions. In everyday social interactions, faces receive most attention (Adams, Albohn, & Kveraga, 2017; Fletcher-Watson et al., 2008), and are thus arguably something that tends to be prioritized. We use faces to predict and interpret socially relevant information about the person behind the face. For example, we try to infer other people's thoughts, feelings, intentions, and even personality from just a single glance at their faces (Todorov, 2017). Faces are weighed so heavily, they influence even political votes (Antonakis & Eubanks, 2017; Olivola & Todorov, 2010a), juridical decisions (Blair, Judd, & Chapleau, 2004; Porter, ten Brinke, & Gustaw, 2010; Wilson & Rule, 2015; Zebrowitz & McDonald, 1991), financial trust (Chang, Doll, van 't Wout, Frank, & Sanfey, 2010; Rezlescu, Duchaine, Olivola, & Chater, 2012), hiring decisions (Hassin & Trope, 2000), and more (Todorov et al., 2015). Research in social person perception has even linked specific facial appearances to predictions of specific, perceived personality traits, such as trustworthiness and dominance (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Todorov et al., 2013), showing how people can be (dis)advantaged in social interactions simply by their natural facial appearance. Importantly, although people largely agree on such face based impressions, consensus does not guarantee accuracy (Antonakis & Eubanks, 2017; Efferson & Vogt, 2013; Todorov, 2017).

Studies on face based impressions and their consequences usually manipulate face stimuli to investigate the influence of facial appearance on social judgments. This gives the impression that we perceive faces similarly to how they appear out there. However, faces are very rich and potentially ambiguous

stimuli in themselves. If our cognitive systems actively construct reality opposed to passively record it, should they not also use available information to predict and interpret the visual aspects of the faces they encounter? Indeed, impressions formed from faces are influenced by both target and perceiver characteristics, as well as their interactions (Hehman, Stolier, Freeman, Flake, & Xie, 2019; Hehman, Sutherland, Flake, & Slepian, 2017), suggesting that it matters for face perception what perceivers bring to the table. Given that specific facial appearances are strongly linked to specific person impressions (e.g. of personality traits, mental and emotional states, and behaviors), could perceivers' impressions about the person's traits, states, or behaviors influence their visual experience of the face as well? If so, the influence should run in both directions, which were coined by Hassin and Trope (2000) as 'reading from faces' (RFF: facial appearance influences person impressions) and 'reading into faces' (RIF: person impressions influence facial appearance impressions).

A lot of research has focused on RFF. As outlined above, faces clearly influence our impressions of many other aspects of a person, whether accurately so or not. It makes sense that RFF is persistent, because face based impressions concern non-visual aspects of a person (such as personality or emotional state) that cannot be directly observed and thus need to be inferred. As explained earlier, these impressions are therefore less likely to be disproven, either because there currently are no other cues available to test the impression or because those cues are ambiguous enough to interpret in line with the impression. Moreover, the experience that other people agree with our face based impressions increases our confidence in them (Todorov, 2017).

But what about RIF? Even if verbal information about a person influences people's expectation of that person's facial appearance, they can subsequently perceive the face and compare their prediction of the facial appearance to the sensory input of the facial stimulus that is actually out there. Unlike non-visual aspects, such as personality traits or mental states, that need to be inferred, faces are visually present and thus can be directly observed. Would people in that case still mentally represent the person's face somewhat differently from the actual face that is out there?

READING INTO FACES

Let me clarify a bit more what is meant with RIF. You may have experienced that we use available information to understand facial appearance. For example, knowing whether someone just failed or succeeded to get a much-wanted promotion, helps us to understand the image of that person crying as someone who is crying sad or happy tears (Hassin et al., 2013). But it may be that we still see the same picture in both cases: a person crying. RIF asks whether we also distort the visual facial appearance in our minds based on the available information. For example, do we mentally represent the winner's crying face as looking more competent than the loser's crying face, even if that difference is not really there? Likewise, do we mentally represent someone's face as looking a bit more feminine or masculine based on the information that this person works as a nurse or a truck driver? In other words, is the picture of the face in our minds influenced by the available verbal information?

There is a broad literature on person impression formation showing how information about a target person influences our impressions of the target with the help of our mental models linking the information to other person information (better known as stereotypes; e.g. Quinn, Macrae, & Bodenhausen, 2007). For example, information that someone is black may trigger evaluations of the person as 'athletic' or 'aggressive' (Devine, 1989). Dependent on perceivers' mental models, a person can thus be (dis)advantaged from the start by this simple piece of information about race. If mental models also relate this information to specific facial appearances and these visual stereotypes bias perceivers' mental representations of the person's face, the (dis)advantage is engrained in people's visual representation of the person.

Vision is considered by many to be our primary and highest valued sense, receiving most attention of all the senses both in everyday conversations (San Roque et al., 2015) and scientific research (Hutmacher, 2019). People highly rely on their vision to decide what is true, reflected in the popular expression "seeing is believing". It may therefore appear to perceivers that their visual representation of a person reflects objective truths about the person, untouched by their own subjective expectations (Quinn et al., 2007). If humans indeed read into faces, it is important to make people realize that at least to some

extent "believing is seeing" as well. This is important not only for those who are disadvantaged by biased mental representations of their faces, but also for situations in which accurate mental representations of faces are crucial, such as in eyewitness procedures.

Evidence from the Literature

It has been demonstrated that verbal information about a person influences people's *expectations* about that person's facial appearance. In line with the premise that people represent knowledge perceptually (Barsalou, 1999), information about group membership has been shown to bias people's expectation of the face in line with stereotypes about the group. For example, faces of individuals believed to be ingroup (vs. outgroup) members are expected to appear amongst others more attractive, intelligent, and trustworthy (Ratner, Dotsch, Wigboldus, van Knippenberg, & Amodio, 2014). More Western vs. Eastern looking faces are expected for immigrants who try to adopt the American host culture vs. maintain their heritage culture (Kunst, Dovidio, & Dotsch, 2017). Another study showed that (non-)welfare recipients are expected to appear amongst others more (less) lazy and incompetent (Brown-Iannuzzi, Dotsch, Cooley, & Payne, 2016). Finally, faces of individuals whose group members are known to perform trustworthy (criminal) behavior are expected as more trustworthy (criminal) looking (Dotsch et al., 2013).

However, these studies concern expectations only. Participants have never seen the person's actual face and therefore the facial appearance is something that needs to be inferred based on the available social information. Much like thoughts, feelings, and traits, the face cannot be directly observed in these studies. Under those circumstances, it is not too surprising that perceptual expectations of the face are biased. After all, having never seen the face, the social information is the only information to base the perceptual expectation on. But what about situations in which people do get to see the person's actual face? Are their mental representations of the face still biased by the social information in that case? Or is the predicted facial appearance overwritten by the actual facial appearance as soon as people get to see the person's face?

According to Dynamic Interactive (DI) theory (Freeman, Stolier, & Brooks, 2020), social information can influence mental representations of the face

via these perceptual expectations, even when people get to see the actual face. The DI theory views social face perception as a process of negotiation between the visual input from the face that is out there and the social cognitive processes going on in the perceiver's mind. While different partially activated interpretations are competing in the process to settle on one final mental representation, these social cognitive processes, such as stereotypes or goals, together with the context can shape the final representation. Neural evidence seems to support this idea, showing that perceptual priors (or expectations) appear to be sent from frontal brain regions to posterior regions to be compared to incoming sensory information and to possibly even shape it in line with the prediction (Brooks & Freeman, 2019; Summerfield & Egner, 2009).

Indeed, many studies have demonstrated how visual information that is quickly processed, such as a person's clothing style, race, gender, or body posture can generate predictions about facial appearances associated with these perceptual cues and subsequently influence the categorization of that person's face, for example on race (Freeman, Penner, Saperstein, Scheutz, & Ambady, 2011), gender (Johnson, Freeman, & Pauker, 2012), and emotional expression (Bijlstra, Holland, Dotsch, Hugenberg, & Wigboldus, 2014; Bijlstra, Holland, & Wigboldus, 2010; Van den Stock, Righart, & de Gelder, 2007). These effects were found mostly when faces looked somewhat ambiguous on the concept of interest (Bijlstra et al., 2014; Freeman et al., 2011; Johnson et al., 2012; Van den Stock et al., 2007), or/and when participants were under time pressure to respond (Bijlstra et al., 2010; Freeman et al., 2011; Johnson et al., 2012), or/and when faces were presented only briefly (Bijlstra et al., 2010; Van den Stock et al., 2007).

Intriguingly, Hassin and Trope (2000) showed that also verbal information about a person can bias the mentally represented facial features when seeing the actual face without any time pressure. In this study, verbal information describing a person's personality as mean or kind-hearted biased ratings of specific facial features. For example, targets described as kind-hearted were rated to have amongst others rounder chins, fuller faces, and shorter ears. This occurred not only for faces that visually appeared ambiguous on the personality trait, but even for faces that visually appeared already relatively extremely mean or kind hearted. Another study found evidence that labelling a face as 'Black' or 'White' influenced perception of the skin's lightness, such that 'White' faces

were judged to be lighter than 'Black' faces, even when their objective lightness was manipulated to be the same (Levin & Banaji, 2006).

Verbal vs. Visual Information

The number of studies on visual vs. verbal information referenced above shows that in comparison to the effect of visual information on mental representations of a seen face, the effect of verbal information has received relatively little attention in the literature. Note that visual information is meant in the sense that the information is visually presented like an image, video, or in real life (e.g. image of someone with grey hair and wrinkles). Verbal information is meant in the sense that the information is presented through words, whether orally or written (e.g. the words that someone is old). The same kind of information can thus be presented either verbally or visually. Although relatively little attention has been devoted to the effect of verbal information, the distinction between verbal and visual information is important to make, because it leads to different consequences theoretically, methodologically, and societally.

Visual information, such as skin color, emotional expressions, hair, and clothing style, is present at the same time as the face. Therefore, it seems likely that this can influence the way the face is mentally represented. After all, humans do not perceive facial features as separate parts. Rather, a face is perceived as a whole (Maurer, Le Grand, & Mondloch, 2002), in a broader visual context. Verbal information is not necessarily present in the visual field while viewing the face and thus need not influence the mental representation of the face. For example, verbal information that someone is unemployed does not need to influence the mental representation of that person's seen face. If it does, this teaches us more about how humans perceive their social surroundings. Their experience of the visual information is then not merely influenced by visual input, but also by verbal information that may concern the visual objects, further increasing our theoretical understanding of social perception as a process integrating both visual input and perceiver's conceptual knowledge (Freeman et al., 2020).

Additionally, the distinction between verbal and visual information has societal relevance. You have limited control on what your appearance looks like. You can adjust your clothing, hairstyle, makeup, and facial expressions, but there is only so much that you can manipulate visually. Hence, there is limited influence

through visual manipulation that you and especially others can exert to change how people mentally represent your face. Verbal information, on the other hand, can be far more easily manipulated, both by yourself and by others (who may have either good or bad intentions). Indeed, conversations between humans in relaxed social settings are dominated by social information about the speaker's own social experiences and relationships or about those of others not currently present (Dunbar, Marriott, & Duncan, 1997). Moreover, in the present era of digital communication, initial contact between people is usually in the form of written online communication (e-mail, Messenger, WhatsApp, online dating, online forum, etc.). Since the worldwide outbreak of COVID-19, to which countries from all over the world responded with obligatory lockdowns (Hale et al., 2021), written online communication has become even more prevalent, both in people's professional and personal lives. Consequently, if not only visual but also verbal information may impact how people mentally represent someone's face, there exists a source of influence that is both prevalent and that everybody (not just the face bearer) can mingle with.

Under Which Circumstances Does RIF (Not) Occur?

Earlier work suggests that it is possible for verbal information to bias the visual mental representation of a face, even after seeing the actual face (Hassin & Trope, 2000; Levin & Banaji, 2006). Yet, even if this is possible, how likely is it really to occur? And under which conditions does it become more or less likely? It appears that researchers have focused on showing that the bias exists, without elaborating on the circumstances that make the bias (dis)appear. Yet, it is relevant to understand under which circumstances verbal information about a target person may bias the perceiver's mental representation of that person's seen face for multiple reasons.

First of all, by improving our understanding of the relative contribution of target appearance and perceiver interpretation on the forming of a mental representation, we improve our understanding of the person impression formation process: how people form impressions of others. Humans are extraordinarily equipped to perceive faces with a major brain network related to face perception (Maurer et al., 2002; Todorov, 2017). Perceiving and interpreting faces is something they practice every day of their lives. Their social interactions – and consequently their (social) lives – heavily rely on adequate interpretation

of face signals. It may therefore seem ineffective to have these signals distorted by their visual system. On the other hand, it may be adaptive to let verbal information about a person bias your mental representation of the person's face. Visually biasing the face may help to remember and therefore predict the person's tendencies better. After all, the pictures in our heads are supposed to be representative of the world to help us predict and interact with it (e.g. Bruner, 1957). For example, mentally representing the perpetrator's face as more untrustworthy looking may help to avoid (trusting) this person in the future. Although your mental representation is not completely veracious (and although this may be detrimental to police investigations), the bias is adaptive as it keeps you safe (Galperin & Haselton, 2013). Indeed, humans have been shown to visually distort non-social objects to ensure their own survival, like perceiving a vertical surface as higher from above than from below, presumably preventing humans from falling from great heights (Jackson & Cormack, 2007).

Second, better understanding under which circumstances people read into faces has societal relevance as well. For instance, eyewitness procedures weigh heavily on the assumption that eyewitnesses accurately remember the perpetrator's face. Having a biased mental representation of the perpetrator's face may not only decrease chances that the perpetrator will be caught, but also increase the risk that an innocent person ends up as a suspect because this person's face resembles the eyewitness' mental representation of the perpetrator's facial appearance (Wagenaar, 1989). Additionally, it may seem unethical to represent the face in a biased way, for example if the person was forced to commit the crime or if the person you saw fleeing the scene turned out not to be the perpetrator at all, but an innocent citizen who saw a chance to escape from the crime scene. Importantly, given that facial appearance influences evaluations of and behavior toward the face bearer (Todorov, 2017), the way we mentally represent someone's face may have a range of social consequences for the face bearer. If we better understand potential influences on how we mentally represent faces, we can take these influences into account in situations involving socially relevant decisions where we know that facial appearance plays an important role, such as eyewitness reports, job interviews, and financial and court decisions.

In sum, better understanding of under which circumstances verbal information influences the mental representation of a seen face has both theoretical

and societal relevance. It will increase our theoretical understanding of the extent and circumstances in which our cognitive systems use both available information about the person already present in the brain as well as visual input from the person's face to create a mental image. Specifically, it may teach us more about determinants of the relative influence of these two sources of information. Furthermore, this understanding can help to raise awareness of the circumstances under which people's visual representations of others may be colored. Hopefully, such awareness can encourage people to take reasonable (pre)caution(s) when making socially consequential decisions about others.

Understanding Under Which Circumstances RIF Occurs: A Bayesian Inspired View

As indicated earlier in this chapter, Bayesian models are successfully used to describe brain functioning and perception in the literature (Clark, 2013; Edwards et al., 2012; Kersten & Yuille, 2003; Kilner, Friston, & Frith, 2007; Mamassian et al., 2002). Therefore, it may be useful to adopt a Bayesian inspired view on social face perception as well. Indeed, although they do not explicitly call it Bayesian, the DI theory that social face perception is influenced by bottom-up visual cues and top-down social cognitive factors (Freeman et al., 2020) is consistent with a Bayesian view on perception.

Interestingly, besides informing *how* perceiver's prior knowledge biases social perception, Bayesian models may also inform *under which circumstances* such biases (dis)appear. A Bayesian inspired view can produce general predictions on the circumstances in which verbal information may interact with sensory input from the actual face to result in a visual mental representation of the face. To see how this would work in the context of RIF, let us return to the Bayesian example of the spider and the dot on the wall and reframe it into a situation of social face perception, namely the imagined crime scene from the introductory paragraph. The probability that the perpetrator's face ends up as untrustworthy looking in your mind or P(untrustworthy looking face | sensory input from perpetrator's actual face) depends not only on the probability that your brain would receive sensory input like that from the perpetrator's actual face if there indeed is an untrustworthy looking face out there or P(sensory input from perpetrator's actual face | untrustworthy looking face), but also on the prior probability that you will be seeing an untrustworthy looking face or P(untrustworthy looking

face), which in your mind is potentially influenced by the warning you received that a robbery was going on. In Bayesian terms, those probabilities are called the *posterior probability*, *likelihood*, and *prior probability* respectively.

We can see from this example how taking a Bayesian inspired view may help us to generate some general predictions concerning the circumstances of RIF. First of all, if verbal information about the perpetrator is to influence the visual mental representation of the seen face, then clearly the verbal information needs to generate an expectation about the visual qualities of the face (General Prediction 1: 'Face Prior Hypothesis'). In this case, the verbal information should generate a prior (or prediction) about the trustworthiness appearance of the perpetrator's face.

Second, it appears that the relative strength of the prior probability and the likelihood determine their relative influence on the posterior probability (General Prediction 2: 'Prior-Likelihood Balance Hypothesis'). In other words, if we want to learn more about the circumstances under which verbal information does (not) bias the mental representation of the seen face, we should focus on the relative strength of the prior belief that the perpetrator's face will look untrustworthy and of the sensory input from the perpetrator's actual face. For example, how strongly you expected the face to appear untrustworthy looking may vary depending on amongst others how reliable you found the warning about the robbery, to what extent you believe robbers to be untrustworthy people, and how strongly you associate the concept of trustworthiness with specific facial features. On the other hand, the strength of the sensory input may vary depending on amongst others how well you could encode the perpetrator's actual face (e.g. depending on whether it was dark in the street, the perpetrator was wearing a face mask or sunglasses or even a balaclava, the perpetrator looked in your direction, and whether you were distracted by anything else in the street).

The two general predictions that arise from taking a Bayesian inspired view can be tested in experiments that systematically manipulate specific instances of these general predictions. As such, adopting a Bayesian inspired view can help to guide social face perception research in the formulation of research questions, hypotheses, and experimental operationalizations. It also serves as a

broader theoretical framework within which findings from social psychological experiments can be organized, providing specific instances of the general rules. Figure 1.1 presents a schematic representation of the two general predictions.

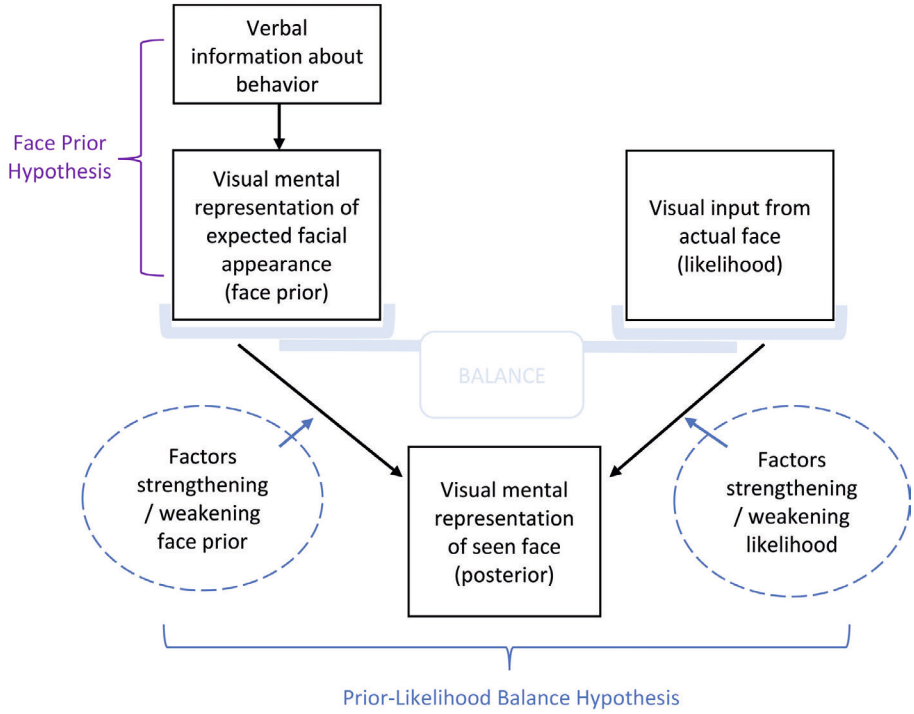


Figure 1.1. Schematic representation of the two general predictions. **Face Prior Hypothesis**: verbal information should generate an expectation about facial appearance (face prior). This could for instance happen because the verbal information activates a person impression (e.g. untrustworthy) that is associated with certain facial features or because the described behaviors are directly associated with certain facial features. **Prior-Likelihood Balance Hypothesis**: the relative strength of the face expectation (face prior) and visual input from the actual face (likelihood) determines their relative influence on the mental representation of the seen face (posterior). The depicted balance scales represent the importance of the relative strength, suggesting that verbal information will only bias the mental representation of the seen face (posterior) if it leads to a face expectation (face prior) that is relatively strong compared to visual input of the actual face (likelihood). Therefore, to understand under which circumstances verbal information does (not) bias the mental representation of the seen face (posterior), we should find out which circumstances strengthen and weaken the face prior and likelihood, depicted here in the blue circles.

HOW TO STUDY MENTAL REPRESENTATIONS: REVERSE CORRELATION METHOD

Although the concept of RIF makes sense theoretically, it poses a methodological challenge for researchers. How can they ever measure participants' mental representations of a person's facial appearance? After all, researchers cannot read people's minds. Indeed, they are unable to visualize people's mental representations directly. However, they can visualize approximations of people's mental representations, learning about the relevant facial features that are crucial to people's mental representations.

So how can they do this? Researchers may ask participants to describe or draw their mental representations. However, participants may be unable to consciously access their mental representations, let alone be able to translate them into words or draw them veraciously. It would be easier for participants if they could recognize facial features from pictures, relying on their gut feeling of a good fit with their mental representation. Researchers could thus present participants with many pictures of faces and see which ones participants select. However, because researchers select the pictures that are presented to participants, they preselect which facial features are (not) manipulated and which are thus presumed (ir)relevant to participants' mental representations.

A data-driven methodology called reverse correlation (RC) overcomes these problems regarding participants' introspective abilities and researchers' assumptions about relevant facial features (Brinkman, Todorov, & Dotsch, 2017; Dotsch & Todorov, 2012; Jack & Schyns, 2017). In a RC task, participants are still presented with many images of faces, so participants can rely on recognition. To overcome the preselection of facial features by researchers, facial features are manipulated randomly.

In a two-images forced choice RC task (Dotsch & Todorov, 2012), researchers select one base image of a face, whose appearance is subsequently altered by superimposing random noise on the image. The noise makes pixels randomly appear lighter or darker, thereby slightly changing the appearance of the facial features. In this way, facial features are manipulated without researchers' assumptions restraining the manipulations. In this RC task, the opposite noise

pattern is also superimposed on the same base face image and participants select the image (out of the two) that corresponds closest to their mental representation of the concept researchers are interested in. For example, participants can be asked to select the more Moroccan-looking face when researchers are interested in their mental representations of a typical Moroccan face (Dotsch, Wigboldus, Langner, & Van Knippenberg, 2008). Participants do this for hundreds of pairs of randomly generated images. See Figure 1.2 for examples of a base face image, a noise pattern, and two images in which a noise pattern and its inverse are superimposed on the base face image.

a b c

Figure 1.2. Examples of a base face image (a), a random noise pattern (b), and a random noise pattern and its inverse superimposed on the base face (c). The base face in this example is the average of the neutral female mean and neutral male mean of the Averaged Karolinska Directed Emotional Faces database (Lundqvist & Litton, 1998).

By averaging all the images a participant selected into a so-called classification image (CI), researchers can visualize which facial features drove this participant's responses, giving an indication of the facial features that are relevant to his or her mental representation. A large variety of mental representations can be investigated in this way even if the base face does not look much like the mental representation. All researchers need to change is the instruction on what concept participants should classify the images. For example, approximations of mental representations of Chinese and Moroccan faces could be visualized using the same Caucasian base face (Dotsch et al., 2008). See Figure 1.3 for the used base face and resulting average Moroccan and Chinese CIs for this study.

a b c

Figure 1.3. The Caucasian base face image (a) and resulting Moroccan (b) and Chinese (c) average CIs from the RC task used by Dotsch and colleagues (2008). Images reprinted with permission.

Usually, researchers are interested not only in visualizing the approximated mental representations as CIs, but also wish to compare the CIs between experimental conditions on a certain concept of interest (e.g. gender, race, emotion, personality). For example, Dotsch and colleagues (2008) wanted to know whether more (vs. less) prejudiced participants mentally visualized Moroccan faces as more criminal and less trustworthy. To this end, the resulting CIs can be scored on the concept of interest (e.g. criminal and trustworthy) by a group of independent raters (i.e. the 'rating method'). These raters did not participate in the original experiment containing the RC task and are therefore ignorant of any manipulations or measures in the experiment itself. They simply rate the CIs on the chosen concept, allowing researchers to subsequently compare the CI scores between conditions.

It is important to note that a resulting CI does not equal the participant's mental representation; it is a visual proxy of that mental representation (Brinkman et al., 2017). Each CI depends not only on the participant's mental representation, but also on the specific base face image chosen by the researchers, the random noise patterns used in the task, the participant's motivation during the task (completing hundreds of trials of noisy images can be experienced as demotivating), and error. That being said, the RC task offers a potent opportunity to learn more about people's mental representations and presently can be considered the best method available for this purpose.

THE PRESENT DISSERTATION

With the present dissertation, we aim to contribute to the understanding of social person perception by focusing on the following question: *Under which circumstances does verbal information about a person's behavior (not) bias the visual mental representation of that person's seen face?* Instead of merely demonstrating whether this bias occurs, we aim to increase understanding of the circumstances under which such a bias is more or less likely to occur. We combine a Bayesian inspired theoretical view with a data-driven RC methodology in our approach to investigating this research question. Below, I will give a brief overview of the dissertation.

Chapter 2 is a methodological chapter that introduces a potentially cost-efficient alternative to the rating method of the RC methodology. Remember that in this method a group of independent raters is asked to score all CIs on a concept of interest. The rating method can be considered cost inefficient when applying sequential hypothesis testing (as can be done with Bayesian statistical analyses) and when a large number of individual CIs needs to be rated (as is the case with large sample sizes). In this chapter, we introduce a **criterion creation method** as an alternative way to score individual CIs on a concept of interest. We demonstrate how to create and use such a criterion and compare its results to the rating method.

Chapter 3 attempts to test the two general predictions derived from the Bayesian inspired view introduced above. Study 3.1 tests the Face Prior Hypothesis by investigating whether **verbal information about a person's behavior** generates an expectation about the facial appearance of that person (called face prior). The behavioral information was manipulated to create the impression that the target person was a trustworthy or untrustworthy person. Studies 3.2-3.4 test the Prior-Likelihood Balance Hypothesis by increasingly shifting the balance in strength between the face prior and the sensory input from the actual face. We varied the strength of sensory input from the actual face through (not) instructing participants to **remember the face well** and through manipulating the **face presentation duration**. We hypothesized that a goal to remember the face well as well as a longer face presentation duration would lead to better encoding of the visual input from the face, which could increase its influence on

the final mental representation. We varied the strength of the face prior through (not) instructing participants to actively **mentally visualize their face prior** before viewing the actual face. We hypothesized that actively visualizing the expected facial appearance would ensure an accessible and detailed face prior, which could then influence the final mental representation more strongly.

Chapter 4 continues testing the Prior-Likelihood Balance Hypothesis through different operationalizations. In both studies, we instructed participants to remind themselves of their impression of the person before starting the RC task. We attempted to weaken information from the actual face by adding a **time delay** between the face presentation and the RC task in one of the two studies. The idea was that it would be harder to remember facial details compared to the trustworthiness impression (induced by the behavioral information) over time. This should weaken the available information from the actual face, while the face prior (reflecting the trustworthiness impression) should remain relatively unchanged, thus increasing the influence of the latter on the final mental representation over time. We investigated this when face presentation duration was relatively long to make it more comparable to most real-life situations. Moreover, we explored whether the effects would be similar for different face identities.

In Chapter 5, we return to the main research question of the present dissertation and attempt to answer the question in light of the research findings presented in the previous chapters. We provide a critical discussion of the present methods and research findings, and consider theoretical, methodological, and societal implications.

A Note on Open Science Practices

Over the last years, the field of psychology has taken notable effort to increase the transparency and quality of its research, aiming to generate so-called 'open science'. These efforts are a result of an impactful replication crisis in the field, during which it discovered that the research findings of many published studies could not be replicated (Open Science Collaboration, 2015). It became clear that scientific publications sometimes omitted important information (e.g. on excluded participants, measures, or analyses) or misreported information (e.g. reporting hypotheses or analyses as though they were predicted beforehand

whereas they were in fact added exploratively after initial results were known). Although exploration is important for a scientific field to advance and should not be discouraged, transparency is necessary to advance reliably (Wigboldus & Dotsch, 2016). Transparency about research practices allows other researchers to distinguish exploratory from confirmatory analyses and to replicate and build on the reported work.

In the present dissertation, I aimed to contribute to open science practices by preregistering all experiments on the Open Science Framework (https://osf.io/) and by reporting my research transparently. The links to the preregistrations are given in the chapters discussing those experiments. I followed the Research Data Management Protocol of the Behavioural Science Institute to ensure secure data management. Moreover, materials, anonymized data, and analyses of all studies are available by contacting the Research Data Officer of the Behavioural Science Institute at Radboud University. I further aimed to improve the quality of my research by collecting data from large and diverse samples from varying parts of the world, and by adopting Bayesian statistical analyses, which allow for sequential hypothesis testing as well as quantification of the amount of evidence for the null model compared to the alternative model (Dienes, 2016; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017).


2

Validation of Automated Scoring of Perceived Facial Trustworthiness in the Reverse Correlation Task

This chapter is based on:
Jansen, L.F., Dotsch, R., Holland, R.W., & Wigboldus, D.H.J. (2021). *Validation of Automated Scoring of Perceived Facial Trustworthiness in the Reverse Correlation Task*. Unpublished manuscript. Radboud University, Behavioural Science Institute, Nijmegen, The Netherlands.

ABSTRACT

As impressions from faces are consequential for behavior and attitudes, scientific interest in people's mental representations of faces has increased. A popular technique to visualize approximations of such mental representations is reverse correlation. An independent group of raters typically rates the resulting visualizations, called classification images (CIs), on a concept of interest. However, this rating method is cost inefficient when researchers wish to conduct sequential hypothesis testing, as well as when rating all individual (opposed to group) CIs, which is necessary to keep Type I error rate under control. A criterion creation method may solve this problem, preventing the need to collect new ratings from a new group of raters before each (sequential hypothesis) test. In the present work, we demonstrate how to create and validate a criterion with which CIs can be scored on a concept of interest, using perceived facial trustworthiness as an example. We demonstrate how to use the criterion in new studies and compare its results to those of traditional ratings. We propose a combination of both methods for the most efficient and optimal test, allowing researchers to benefit from the advantages of both the reverse correlation task and sequential hypothesis testing.

Keywords: reverse correlation, classification image, criterion, Sequential Bayes Factors, projection, perceived facial trustworthiness

Knowing that face-based impressions have a significant impact on attitudes and behavior (Todorov, 2017), researchers in social person perception have become increasingly interested in how people perceive faces. Does Donald Trump's face look more competent in the mind of a Trump voter opposed to a Biden voter? Does a stranger's face look more untrustworthy in the mind of a police officer when estimated to be the perpetrator rather than the victim? Questions like these concern pictures of faces in people's minds (Lippmann, 1922), also called their mental representations of these faces. Although scientists still cannot read minds, one increasingly used method, called reverse correlation (Brinkman, Todorov, & Dotsch, 2017; Dotsch & Todorov, 2012; Jack & Schyns, 2017), can visualize *approximations* of such mental representations.

Reverse Correlation

Reverse correlation (RC) is increasingly applied in the field of social person perception (e.g. Brown-Iannuzzi, Dotsch, Cooley, & Payne, 2016; Dotsch & Todorov, 2012; Dotsch, Wigboldus, & Van Knippenberg, 2013; Ratner, Dotsch, Wigboldus, van Knippenberg, & Amodio, 2014). Although multiple variations of RC tasks exist, the RC task that appears highly popular in social face perception research is the noise-based two-images forced choice RC task (Dotsch & Todorov, 2012), first used by Dotsch, Wigboldus, Langner, and Van Knippenberg (2008). In this task, random stimuli are created by superimposing random noise patterns (i.e. random grayscale pixel values) on the same base image of a face, slightly altering the appearance of the image. On each trial, two images are presented. Both images consist of the base face image and a unique random noise pattern, one pattern being the inverse of the other. On each trial, participants select the one image (out of the two) that looks most similar to their mental representation.

By averaging the noise patterns of all the images a participant selected, an individual classification pattern (CP) is generated for this participant. By superimposing the individual CP on the base face image, the classification image (CI) for this participant is visualized, called an individual CI. The individual CI is interpreted as an approximation of that participant's mental representation (Brinkman et al., 2017; Dotsch & Todorov, 2012). All individual CPs (one per participant) can be averaged into group CPs to arrive at that group's average classification pattern. This group CP, too, can be visualized into a group CI by

superimposing it on the base face (Brinkman et al., 2017). Summarized in an equation, classification pattern (CP) + base face = classification image (CI).

Once researchers have obtained the relevant CIs, they usually want to have them scored on a concept of interest. This concept can be anything, such as a personality trait, emotional expression, mental state, race, age, gender, attractiveness, weariness, gaze direction, and so on. How does the face come across on the chosen concept? To get these scores, researchers typically ask an independent group of participants (*raters*) to rate each CI on the concept of interest (e.g. "how trustworthy does this face look?"). After standardizing ratings per rater, the average rating across raters per CI serves as the score for this CI.

This is where a problem comes in. The rating method works well after all CIs have been collected. However, it is inefficient when data need to be analyzed during the data collection process. For example, a statistical method gaining popularity among social psychologists is Bayesian inference (van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017). It allows researchers to monitor the data as they come in through sequential hypothesis testing with Bayes factors, or Sequential Bayes Factors (SBF; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017). Consequently, researchers can stop data collection once the Bayes factor has reached a predetermined threshold (e.g. when there is strong evidence for the null or alternative model; e.g. $1/10 \geqslant BF_{10} \geqslant 10$). This potentially saves a lot of time and resources. Yet, it also means that researchers wish to analyze their data after each new batch. Using the rating method, researchers have to collect data from a new group of raters each time they wish to analyze their data. This quickly takes up a lot of time and resources. Therefore, the rating method is problematic if researchers want to analyze the data as they come in.

The rating method is also problematic when researchers collect ratings of group CIs only, which is common practice in reverse correlation literature (Jeremy Cone, Brown-Iannuzzi, Lei, & Dotsch, 2020). It is understandable that researchers often opted for rating only group CIs, because their signal-to-noise ratio is better than that of individual CIs, and because it seems more efficient to have raters rate a couple of group CIs opposed to hundreds of individual

CIs. However, using ratings of only group CIs has been shown to inflate Type I error rate, which does not occur for ratings of individual CIs (Jeremy Cone et al., 2020). Now that rating only group CIs is no longer acceptable and having raters rate all individual CIs is still quite costly, the quest for efficient scoring methods of individual CIs has started.

Criterion Creation Method

A solution to both problems outlined above could be to create a criterion CI for the concept of interest to which CIs can be compared. To create a criterion CI for the concept of interest, only one group of participants (call them *criterion creators*) is needed to generate CIs that represent their mental representations of the concept of interest. This idea was derived from Imhoff and Dotsch (2013), who investigated how similar German participants' mental representations of Europeans were to their mental representations of Germans and of themselves. Participants completed three RC tasks, one for each category (European, German, and self). For each participant, the pixel luminance values of the resulting European CP were correlated to those of the German and self-CP to determine whether participants used ingroup-projection, self-projection or both when mentally representing members of a superordinate group. Similarly, researchers can ask one group of criterion creators to generate CIs that represent their mental representation of the concept of interest, conduct their study, and score each participant's CP from the study on the concept of interest by correlating it with the average criterion CP (also suggested by Brinkman et al., 2017). This way, researchers can easily score the CIs in the study and do not need to collect new ratings each time new data have come in. Once a valid criterion CI is created, researchers can reuse it to efficiently score individual CIs of new studies as well.

Instead of calculating the correlation between a participant's CP and the criterion, we propose to use the projection of the participant's CP on the criterion as a measure of similarity. Although correlations and projections lead to highly similar results, projections have the advantage that they capture slightly more information. To explain this, imagine the participant's and criterion CPs as two points in a multidimensional space. For each CP, imagine an arrow from the origin of the space to the position of the CP, called that CP's vector. Whereas correlations only focus on the angle between the vectors of

the two CPs, projections also consider the length of the participant's CP vector (see Fig. 2.1).

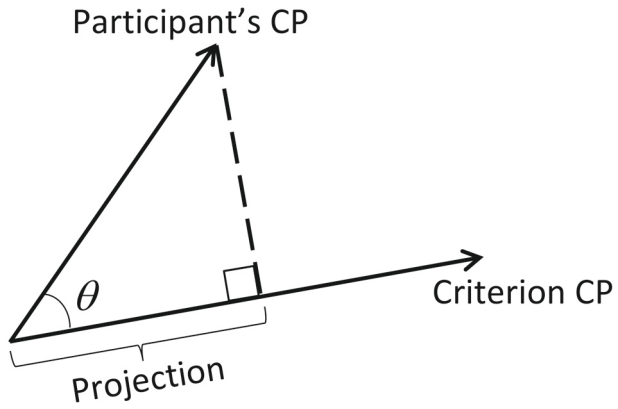


Figure 2.1. Whereas a correlation (= cosθ) focuses on the angle between the participant's CP vector and the criterion CP vector, the projection takes into account both the angle and the length of the participant's CP vector.

The Present Research

The criterion method thus allows researchers to benefit from both the RC task and the option to analyze data as they come in (as with SBF). If the method appears valid, researchers could even use the criterion CI for all future studies in which CIs need to be scored on the concept captured by the criterion. That is, if the criterion creators are representative for the participants recruited in the future studies. In that case, the method would eliminate the need for an extra data collection from a group of raters, making future studies even more efficient.

In the current study, we aim to demonstrate how researchers can create and test the validity of a criterion CI. Moreover, we aim to demonstrate how to use the criterion CI to score CIs from new studies and how its results compare to those of the rating method. For this demonstration, we focus on perceived trustworthiness as the concept of interest. The result is a perceived trustworthiness criterion CI, visualizing what needs to change in a face to go from untrustworthy to trustworthy looking. We chose perceived trustworthiness as it is one of the major underlying dimensions of person impression formation (Todorov, Said, Engell, & Oosterhof, 2008), with significant consequences for subsequent behavior and attitudes towards the evaluated individual (Porter et al., 2010).

METHOD

The creation of the criterion CI was preregistered on the Open Science Framework (https://osf.io/yxzf5).

Participants

The criterion CI is built on the responses of 100 Caucasian adults (42 women, 58 men, M_{age} = 32.46, SD_{age} = 10.72) with normal or corrected-to-normal vision, situated in varying countries across the world (see Table 2.1 for an overview) with education backgrounds varying from primary education to doctoral or equivalent level (see Table 2.2 for an overview). Participants received £2.50 for participation. We collected data on Prolific Academic (https://www.prolific.ac) from 116 participants. Sixteen participants were excluded from analyses following our preregistered exclusion criteria. Nine participants were excluded because their median reaction time on the RC task was less than 500 milliseconds. Seven more participants were excluded because they failed to follow the instruction to use each response option in the RC task at least once.

Table 2.1. Number of participants per country. The information is based on participants' reported current country of residence.

Country	Number of participants
Australia	1
Austria	1
Bosnia and Herzegovina	9
Canada	3
Croatia	1
Czech Republic	1
Georgia	3
Greece	1
Hungary	2
Italy	2
Lithuania	2
Mexico	1
Russia	2
Serbia	6
South Africa	1
Spain	3
Sweden	1
United Kingdom (UK)	43
United States of America (USA)	16
Unknown	1
Total	**100**

Table 2.2. Number of participants per highest completed education level.

Highest completed education	Number of participants
Primary education	1
Lower secondary education	3
Upper secondary education	20
Post-secondary non-tertiary education	9
Short-cycle tertiary education	3
Bachelor's or equivalent level	47
Master's or equivalent level	16
Doctoral or equivalent level	1
Total	**100**

Procedure

Participants completed the study online. Before starting the experiment, participants provided informed consent, were asked to ensure a quiet environment without distractions, and set the window to full screen.

Participants completed a four-alternatives forced choice RC task in which they rated 500 randomly generated faces on trustworthiness. In this RC task, participants view one image per trial presented with 4 response alternatives. The response options were 'probably untrustworthy', 'possibly untrustworthy', 'possibly trustworthy', and 'probably trustworthy' (advocated by Mangini & Biederman, 2004; Murray, Bennett, & Sekuler, 2002). Participants were asked to use each response option at least once. With 4 response options, this RC task allows CIs for both perceived untrustworthiness and perceived trustworthiness to be created for the same sample of participants. This enables the visualization of what needs to change in a face to go from untrustworthy to trustworthy looking. Participants could take a break after every 100 trials if they wanted to.

All stimuli consisted of the same base face image, which is the grayscale average of the average male and average female faces of the Karolinska faces database (Lundqvist, Flykt, & Öhman, 1998), and different random noise patterns superimposed on the image. The noise was composed in the same way as described by Dotsch and Todorov (2012). All images were sized 512 x 512 pixels. See Figure 2.2 for the base face and an example of a random noise pattern

superimposed on the base face. On each trial, the stimulus was randomly drawn without replacement from the pool of stimuli. After the RC task, participants filled out a short questionnaire and were redirected to Prolific.

Figure 2.2. Base face used in the reverse correlation task (left), and a random noise pattern superimposed on the base face (right).

RESULTS

Criterion CI Creation

Underlying CIs

The criterion CI is based on the faces that were categorized into the extreme response options (‘probably untrustworthy’ and ‘probably trustworthy’). To generate the CIs for each of these response options per participant, first the noise patterns that were classified under that response label were averaged into individual CPs, one CP per participant. This was done by averaging the parameters on which those noise patterns are based, resulting in 4092 mean parameters per participant per response label (Dotsch & Todorov, 2012). By superimposing these individual CPs on the base face, the individual CIs for each response option were generated. In order to indicate what the CIs for each response option across participants should look like, the 4092 mean parameters of each participant (the individual CPs) were averaged across participants (into group CPs). By superimposing these group CPs on the base face, the group CIs for each of the two high confidence response options were generated,

resulting in a group perceived untrustworthiness CI and a group perceived trustworthiness CI.

Perceived trustworthiness criterion CI

To create the perceived trustworthiness criterion CI, we applied the procedure of Mangini and Biederman (2004). The group perceived untrustworthiness CP was *subtracted* from the group perceived trustworthiness CP, resulting in the perceived trustworthiness criterion CP. Naturally, the group CPs belonging to the untrustworthy and trustworthy condition will show some overlap. After all, not every single pixel of the face and background needs to change to create a difference relevant to trustworthiness appearance. By using subtraction, the criterion CP zooms in on those signals that differentiate a more trustworthy appearing face from a more untrustworthy appearing face (Mangini & Biederman, 2004). In order to visualize the criterion CP into the criterion CI, this CP (+1) and its inverse (-1) were superimposed on the base face, resulting in visualizations of a perceived trustworthiness (+1) and perceived untrustworthiness (-1) criterion face. See Figure 2.3 for these visualizations. Notice that by multiplying the criterion CP with increasingly positive (negative) numbers and superimposing each on the base face, increasingly trustworthy (untrustworthy) appearing criterion faces can be visualized. This gives the criterion CI the characteristic of a perceived facial trustworthiness dimension.

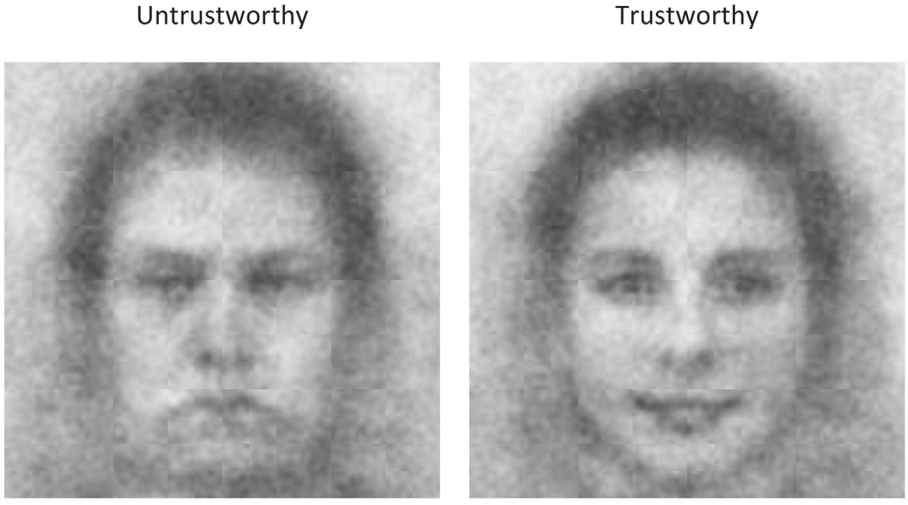


Figure 2.3. Visualization of the perceived trustworthiness criterion CI (right image) and its inverse (left image), forming two points of a perceived facial trustworthiness dimension.

Criterion CI Validation

Face validity

Inspecting the criterion CI and its inverse on face validity, we conclude that they shows a clear variation on trustworthiness similar to earlier visualizations of perceived facial trustworthiness in the literature (Dotsch & Todorov, 2012; Todorov, Dotsch, Porter, Oosterhof, & Falvello, 2013). Interestingly, the importance of masculine and feminine facial features comes across clearer in our visualization, with perceived untrustworthy faces seemingly showing more masculine facial features and perceived trustworthy faces seemingly showing more feminine facial features. The basis for our criterion CI was a gender-neutral face, opposed to male faces used in studies by Dotsch and Todorov (2012) and Todorov and colleagues (2013), possibly allowing the influence of gender to become more articulated.

Untrustworthiness vs. trustworthiness

We based the criterion CI on the responses of 100 participants with varying backgrounds, because we want the criterion CI to be representative for a large group of people. Although people may show differences in their mental representations of facial trustworthiness, they also show much agreement (Todorov, 2017). Creating a criterion CI that captures the difference between perceived facial untrustworthiness and trustworthiness would make no sense if there is no consensus whatsoever between participants on what constitutes an (un)trustworthy appearing face or if participants believe untrustworthy and trustworthy faces look about the same. Therefore, we wanted to check whether individual perceived untrustworthiness CPs generally differed from individual perceived trustworthiness CPs.

To this end, we computed Euclidean distances on a matrix holding the individual perceived untrustworthiness and trustworthiness CPs. This creates an impression of how dissimilar each CP is to the other CPs in the dataset. To understand this, imagine each CP as a point in a multidimensional space, with the distance between the points representing the similarity between the CPs. The further apart the points, the less similar they are. To visualize the Euclidean distances in a plot, they were submitted to a multidimensional scaling analysis with two dimensions (see Fig. 2.4 for the result). The plot showed two

extreme outliers, which pressed all remaining data points into one small cluster, making it hard to interpret the plot (see Fig. 2.4A). The Euclidean distances were therefore computed without these two outliers in the dataset and submitted to a multidimensional scaling analysis like before. The X-axis now showed a clear distinction between perceived untrustworthiness and trustworthiness CPs, suggesting that individual perceived untrustworthiness and trustworthiness CPs indeed generally differed from each other. The Y-axis seemed to add little information.

This observation is confirmed by a logistic regression predicting trait (perceived untrustworthiness / perceived trustworthiness) of the individual CPs from the coordinates on the X-axis and Y-axis. If an individual CP's coordinate increases by one unit on the X-axis, the odds that this CP portrays perceived trustworthiness (opposed to untrustworthiness) increase by approximately 22.41, 95% CI [9.94; 58.84], opposed to only 0.09, 95 % CI [0.03; 0.27], for a one unit increase on the Y-axis.

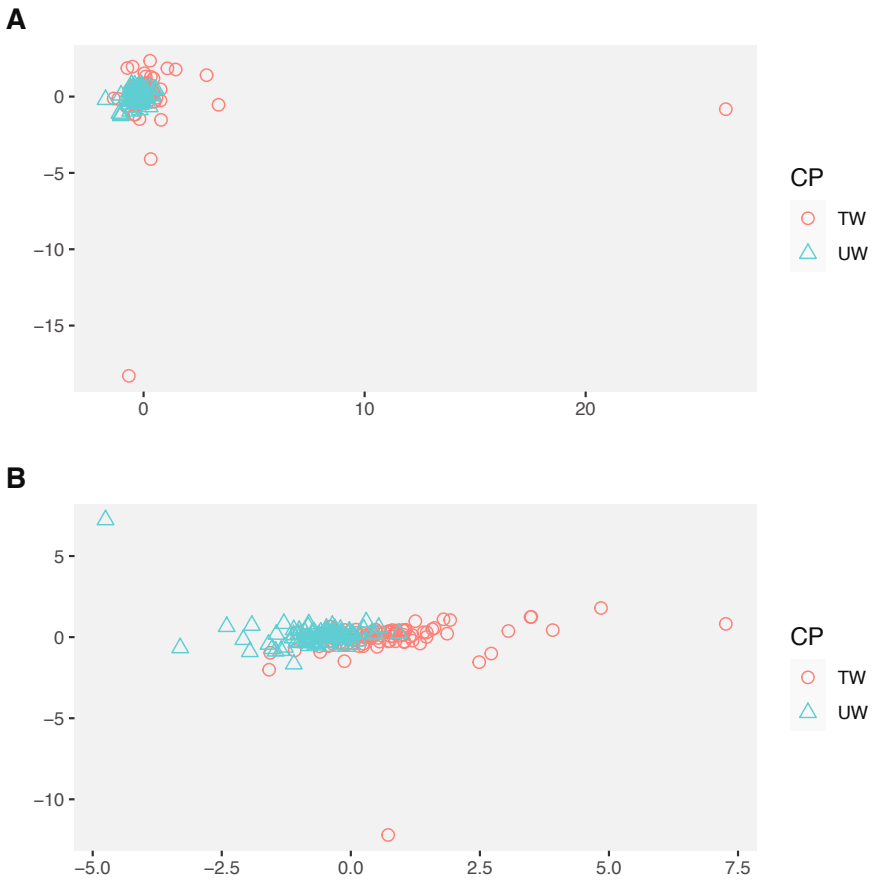


Figure 2.4. Multidimensional scaling solution performed on the Euclidean distances of the individual CPs for the perceived untrustworthiness (UW; green triangles) and trustworthiness (TW; red circles) categories. The original solution was hard to interpret due to two outliers (A). The solution was therefore recomputed on the dataset without these two outliers, showing the X-axis as a discriminator between perceived untrustworthiness and trustworthiness CPs (B).

Predictive validity

The criterion CI is created to compute projection values for new CIs, indicating how much perceived trustworthiness each CI contains according to the perceived trustworthiness criterion CI. The next step is to validate the projection value provided by the criterion CI. For each individual CI used to create the criterion CI, we know whether the participant was thinking of an untrustworthy or trustworthy face, because they categorized the underlying noise patterns as 'probably untrustworthy' or 'probably trustworthy'. Therefore, we can use

these individual CIs to determine whether projection values computed on the criterion CI can predict whether participants aimed to create an untrustworthy or trustworthy looking face. To this end, we can conduct a logistic regression predicting trait (perceived untrustworthiness / perceived trustworthiness) by projection values.

However, since these individual CIs were used to build the criterion CI, this validation is circular. In order to see if the projection values of *new* CIs can predict whether these CIs were supposed to portray perceived facial untrustworthiness or trustworthiness, we performed a cross validation on the data. We computed the criterion CI based on the CIs of the first half of participants only (Criterion1) and computed the projection values for the CIs of the second half of participants only (CIs2), and vice versa (Criterion2 and CIs1).

Projection values

Projection values were computed as follows. We first converted both the 2D individual CP and the CP of the perceived trustworthiness criterion CI to 1D vectors. In other words, we converted each 2D matrix holding the pixel values into one long row, or a 1D vector. We then computed the projection by taking the dot product of the individual CP and the perceived trustworthiness criterion CP, divided by the sum of squares of the components of the perceived trustworthiness criterion CP.[1] This computation returns a single value, which we interpret as the perceived trustworthiness score for the individual CI. It quantifies how much of the perceived trustworthiness relevant cues captured in the criterion CI is present in the individual CI. In other words, it indicates the extent of perceived facial trustworthiness that can be found in the individual CI.

1 This computation corresponds with the highlighted part of the vector projection formula: *vector projection* = $\frac{u \cdot v}{v \cdot v} v$, where u = the participant's CI and v = the criterion CI. It returns the value with which the criterion CI needs to be multiplied to arrive at the projection vector, conveniently providing us with a single value opposed to a vector as trustworthiness score. A more common calculation might have been the scalar projection: *scalar projection* = $\frac{u \cdot v}{|v|}$, which also returns a single value and corresponds to the length of the projection vector. Although the absolute values resulting from the two methods lie on different scales, our computation and the scalar projection should lead to highly similar results. This is confirmed by our simulations based on 1000 iterations showing that the results of the two computation methods correlate r = .997.

Logistic regression

The logistic regression predicting trait (perceived untrustworthiness / perceived trustworthiness) of the individual CIs of the first half of the participants (CIs1) based on their projection values on the criterion CI created on the data of the second half of participants (Criterion2) yields a coefficient for projection values of 13.31 and an odds ratio of 6.05×10^5, 95% CI [3580.69; 1.49×10^9]. For these new CIs, if a projection value on this criterion CI increases by one unit, the odds that the CI is supposed to portray perceived facial trustworthiness increase by approximately 6.05×10^5.

The logistic regression predicting trait (perceived untrustworthiness / perceived trustworthiness) of the individual CIs of the second half of the participants (CIs2) based on their projection values on the criterion CI created on the data of the first half of participants (Criterion1) provides a coefficient for projection values of 6.66 and an odds ratio of 783.94, 95% CI [73.65; 18117.04]. For these new CIs, if a projection value on this criterion CI increases by one unit, the odds that the CI is supposed to portray perceived facial trustworthiness increase by approximately 783.94. Thus, even when the individual CIs were not used to build the criterion CI, the projection values were good predictors of which trait the participant had in mind when constructing the individual CI.[2]

Together, the findings on these validation methods suggest that the perceived trustworthiness criterion CI forms a valid criterion to efficiently score CIs on perceived facial trustworthiness. In the next section, we demonstrate how the criterion CI can be used in new studies and how its results compare to trustworthiness ratings provided by human raters.

2 Dividing the data into two different random halves leads to similar results (e.g. coefficient = 6.65, odds ratio = 772.85, 95% CI [74.15; 16161,64] and coefficient = 30.83, odds ratio = 2.44×10^{13}, 95% CI [9.44×10^6; 5.83×10^{25}]).

Using the Criterion CI in New Studies

Data collection stopping rule

We used the perceived trustworthiness criterion CI to provide us with a stopping rule during data collection of four new experiments, namely those described in Chapter 3.[3] We used the criterion to score participants' CIs on perceived trustworthiness and used these scores as dependent variable in our Bayesian analyses after each new batch of data was collected. We preregistered to continue data collection until the Bayes factor was at least 10 (regardless of whether this was in favor of the tested or the null model) or until we reached a predetermined number of participants. Once we had stopped data collection, we would also use the rater method to score each CI on perceived trustworthiness by having a new group of human raters rate all CIs on trustworthiness.

Comparing results

To illustrate how projection-based trustworthiness scores provided by the criterion CI compare to trustworthiness ratings provided by human raters, we present the correlations between the scores and ratings for each study. We computed the correlations using a Bayesian correlation analysis with default prior settings (a stretched beta prior distribution with width = 1, i.e. a uniform distribution) in JASP (JASP Team, 2020). Table 2.3 presents the estimated correlations with 95% credible intervals between brackets.[4] For each correlation, BF_{10} indicates the amount of evidence for the alternative hypothesis that the correlation is not zero compared to the null hypothesis that it is zero. The estimated correlations range from $r = .39$ to $r = .87$. Although the criterion CI did not deliver the exact same scores as human ratings, the positive correlations suggest that the scores can serve as indicators during data collection.

3 The RC task in the studies of Chapter 4 used a different base face than the ones of Chapter 2 and 3. For the projection values to be meaningful, the base face used to create the criterion CI needs to be the same as the base face used to create the new CIs. Therefore, the criterion CI was not used in Chapter 4.

4 The 95% credible interval indicates a 95% probability that the true value lies between the lower and upper values of the credible interval, assuming that the alternative hypothesis is true.

Table 2.3. Correlations between projection-based trustworthiness scores and human trustworthiness ratings for Studies 3.1-3.4 of Chapter 3.

Study	Correlation [95% credible interval]	Bayes factor (BF_{10})
3.1	r = .87 [.76; .93]	4.24×10^{10}
3.2	r = .51 [.39; .61]	9.31×10^{9}
3.3	r = .39 [.26; .49]	5.93×10^{5}
3.4	r = .49 [.37; .58]	2.29×10^{10}

For further comparison, we present the results of analyses with the projection-based scores vs. the ratings as dependent variable and indicate whether they led to similar or different conclusions for each study. See Table 2.4 for the results.

Table 2.4. Comparing conclusions based on the results as provided by projection-based trustworthiness scores and human trustworthiness ratings for Studies 3.1-3.4 of Chapter 3.

Study	Conclusions	Projection-based results	Rater-based results
3.1	Similar	δ = -2.13 [-2.96; -1.32], $BF_{-0} = 1.15 \times 10^{6}$	δ = -3.05 [-4.01; -2.12], $BF_{-0} = 3.34 \times 10^{9}$
3.2	Similar	δ = .01 [-.28; .30], BF_{01} = 6.02	δ = -.02 [-.31; .27], BF_{01} = 5.95
3.3	Slightly different	δ = -.24 [-.51; .03], BF_{01} = 1.49	δ = -.16 [-.43; .11], BF_{01} = 3.25
3.4	Somewhat different	δ = -.29 [-.56; -.02], BF_{10} = 1.36	δ = -.38 [-.66; -.11], BF_{10} = 6.52

The comparisons of results show that using the criterion CI to score new CIs on perceived trustworthiness led to highly similar conclusions as when asking actual human raters to rate the new CIs in 2 out of 4 studies. In Study 3.3 and 3.4, results from both methods were in the same direction, but human trustworthiness ratings provided somewhat stronger evidence (whether for the null or the alternative model) than projection-based scores.

Sensitivity to noise in CIs

Because some raters in Study 3.2 mentioned that their trustworthiness evaluations were likely influenced not only by the facial characteristics but also by the noisiness of the CIs, we correlated CI trustworthiness with CI picture quality in Study 3.2. Due to deviations from normality and outliers we

report the correlation coefficient Kendall's tau (r_τ).[5] The standardized average noisiness of individual CIs (as judged by 3 independent raters; rater's consistency scores varying between ICC(3,1) = .49 and .75) correlated negatively with CI trustworthiness ratings (r_τ = -.26 [-.36; -.16], BF_{10} = 44444.67). The noisier the CI, the less trustworthy raters evaluated the CI. Projection-based trustworthiness scores, however, were not influenced by noisiness of individual CIs (r_τ = .01 [-.10; .11], BF_{01} = 9.92). We discuss the implications of these findings in the discussion.

DISCUSSION

In the present study, we created and validated a perceived trustworthiness criterion CI with which new CI's can be scored on perceived facial trustworthiness appearance. We propose the criterion creation method as a solution for researchers who wish to both (a) use the RC task in their studies and (b) analyze the data as they come in, for example with SBF. The criterion creation method prevents having to collect new ratings every time additional data are collected. Once the criterion CI is created, it allows researchers to score new CIs with one simple computation, saving both time and resources. Given this cost efficiency, we also explore the suitability of the criterion CI as a full replacement for individual CI ratings. With our demonstration, we hope to equip researchers with the means to create and validate their own criterion CI for the concept of their interest.

Criterion CI Validity

We believe our perceived trustworthiness criterion CI has proven valid in a number of ways. First, it conceptually replicated earlier visualizations of perceived facial trustworthiness in the literature (Dotsch & Todorov, 2012; Todorov et al., 2013). By using a gender-neutral base face, it even seemed to allow the effect of gender cues on facial trustworthiness appearance to emerge (Sutherland et al., 2013; Todorov, 2017). Second, projections of new individual CIs on the criterion CI accurately predicted whether the participant making the individual CI classified it as untrustworthy or trustworthy. Third, projection-based trustworthiness scores of individual CIs correlated positively with trustworthiness ratings of individual CIs provided by actual human raters (in Chapter 3). Even the conclusions of analyses using either one as dependent

5 Correlation coefficients Kendall's tau and Pearson's rho lead to similar conclusions here.

variable were highly similar, though somewhat more modest for the projection-based trustworthiness scores. These results substantiate the idea that the perceived trustworthiness scores provided by the criterion CI can serve as valid indicators of actual human ratings.

Some additional strengths of the perceived trustworthiness criterion CI are worth mentioning. First, our method shows that the conceptual replication of earlier findings on perceived facial trustworthiness can also be found with a four-alternatives forced choice RC task, not only with a two-images forced choice RC task (Dotsch & Todorov, 2012). The four-alternatives forced choice RC task has the advantage that it allows participants to indicate their confidence in their decisions and to leave out noise patterns from the CP that were relatively uninformative regarding the judgment of interest (Brinkman et al., 2017).

Second, the criterion CI is based on the perceptions of 100 Caucasian participants from varying parts of the world with varying educational backgrounds. This is a decently large and variable group to be representative for a large amount of people, making the criterion more representative than the results of most other lab studies with Caucasian participants. We chose to select Caucasian participants with the setup of our upcoming studies in mind (see Chapter 3). Naturally, this does raise the question whether the criterion is also representative for non-Caucasian people. Of course, researchers can make their own choices regarding the sample of participants when creating their own criterion CI, allowing it to be as representative as they wish. Moreover, the same consideration holds for the rating method. In that case, the diversity in the sample of raters determines the representativeness of the ratings.

Together, these observations suggest that the perceived trustworthiness criterion CI is indeed a valid and representative criterion to efficiently score CIs on perceived facial trustworthiness, at least when investigating the perceptions of Caucasians, as in our studies. We chose to use perceived facial trustworthiness as example in our demonstration of the criterion creation method. Future studies should clarify whether the criterion creation method works equally well for other facial dimensions. With the current paper as demonstration, we hope to have equipped other researchers to create and validate criterion CIs for their own concept of interest.

Criterion Creation Method vs. Rating Method

A number of differences between the criterion creation method and rating method needs to be addressed.

Sensitivity to perceived facial trustworthiness cues

In order to evaluate new CIs on perceived facial trustworthiness, each method needs to be sensitive to perceived facial trustworthiness cues present in the CI. Although the perceived trustworthiness criterion CI nicely replicates earlier findings on perceived facial trustworthiness, similar to those earlier findings, it does not cover all relevant facial features for perceived trustworthiness and provides therefore a somewhat limited representation of perceived facial trustworthiness. There are multiple reasons for this limitation. First of all, we computed a linear criterion CI, which means that we miss out on non-linear contributions to perceived trustworthiness, such as face typicality (Sofer, Dotsch, Wigboldus, & Todorov, 2015). Human raters on the other hand are able to pick up on all kinds of perceived trustworthiness cues, including non-linear ones.

Second, the criterion CI is limited by the base face and the number and type of noise patterns used in the RC task. The base face is a Caucasian gender-neutral face and the resulting CIs can only divert from the base face by changing the luminance of the pixels in the image. Moreover, which pixels can be changed into which direction depends on the exact noise patterns used in the task. In the present study we used quite a large amount of noise patterns though, namely 500. Earlier studies have shown that participants are able to create CIs that appear Moroccan or Chinese from a Caucasian base face with only 390 noise patterns (Dotsch, Wigboldus, Langner, & Van Knippenberg, 2008). Yet, the criterion CI cannot capture all facial variations relevant to perceived trustworthiness that a human being can recognize.

Type of new CIs that can be scored/rated

Human raters can rate any type of new CIs on the concept of interest. The criterion CI, on the other hand, cannot. It can only score new CIs created with a RC task using the exact same base face image as the one used in the RC task to create the criterion CI. This is because projections are calculated using the CPs (the selected noise patterns). The CPs indicate how the base face should

be changed (to what extent which pixels should become darker or lighter). The base face itself is not encoded in the CPs. Superimposing a CP on a different base face results in an entirely different facial appearance than was intended by the participant. Therefore, both RC tasks need to use the same base face for their CPs to be meaningfully comparable.

Noise sensitivity

The criterion CI was not biased by the quality, or noisiness, of the individual CIs. Because it is sensitive to only specific variations in perceived trustworthiness, it misses out on other variations (such as face typicality), including apparently the noisiness of the CI. As discussed above, human raters are sensitive to larger variations in perceived trustworthiness than the criterion CI. However, they are also more sensitive to the noisiness of CIs. The noisier the image, the less trustworthy it appears to them. Though it is conceivable that participants included noisiness as an untrustworthiness cue in their CIs, sloppy participation (e.g. replying randomly) in the RC task also leads to noisier CIs (due to less signal). In that case, the noisiness is not meant as a signal for perceived untrustworthiness, but is rather meaningless, obscuring meaningful signal in the CIs. Under those circumstances, raters ideally should not be affected by the noisiness of the CIs. Yet, they are. Interestingly, this suggests that it is worth looking at human ratings of group CIs (the average of individual CIs for an experimental group) in addition to those of individual CIs, as noisiness of individual CIs is largely cancelled out in group CIs, improving their signal-to-noise ratio. We explicitly state 'in addition to' because focusing only on ratings of group CIs likely inflates Type I error, which does not occur for ratings of individual CIs (Jeremy Cone et al., 2020).

Suggestion on the Use of Criterion CIs

Taking all of the above into account, we advise researchers to use a criterion CI to cut back on costs when analyzing data as they come in (as with SBF) and in the end have a group of human raters evaluate the final CIs on the concept of interest. The criterion CI leads to scores similar enough to ratings to serve as a good indicator of the ratings and it prevents researchers from having to pay and collect data from a new group of raters each time data should be analyzed. The raters, however, provide a more complete test of concept-relevant cues present in the CIs than the criterion CI can. Although researchers can choose to replace

ratings with criterion CI scores altogether, they should be aware that the gain in efficiency comes at the cost of a more limited test of the concept of interest.

In more detail, we advise to use the criterion CI as follows. After collecting data from a preregistered minimal number of participants, analyze the data with the projection-based scores provided by the criterion CI as dependent variable. Continue doing this after every new batch of participants has been added. Stop data collection when the results have reached a preregistered threshold (e.g. when the Bayes factor is 10 or higher) or when resources have run out. This saves the hassle of collecting new ratings of the CIs after every new batch of participants. However, given that the criterion CI does not capture all concept-relevant cues, this has been a rather specific test of the concept of interest. It is conceivable that CIs differ on concept-relevant cues that were not captured by the criterion CI but that human observers do pick up on. Therefore, irrespective of the results with the projection-based scores as dependent variable, we advise researchers to now collect ratings of the CIs from an independent group of raters and to use these as the dependent variable in the analysis. In the event that the data are inconclusive, collect additional data as well as new ratings and continue this procedure until the data are convincing or until resources have run out.

This way, we believe researchers will have the most optimal test of the concept-relevant cues in CIs while minimizing the loss of time and resources. As such, the criterion creation method enables a more efficient research procedure, allowing researchers to benefit from both the RC task and sequential hypothesis testing during data collection.


3

Under Which Circumstances Does Non-Visual Behavioral Information (Not) Influence Visual Mental Representations of Seen Faces?

This chapter is based on:
Jansen, L.F., Holland, R.W., Dotsch, R., Brinkman, L., & Wigboldus, D.H.J. (2021). *Under Which Circumstances Does Non-Visual Behavioral Information (Not) Influence Visual Mental Representations of Seen Faces?* Unpublished manuscript. Radboud University, Behavioural Science Institute, Nijmegen, The Netherlands.

Part of the research in this chapter was presented in poster format at the General Meeting of the European Association of Social Psychology in Granada on 6 July 2017, and as an oral presentation at the Annual ASPO Conference of the Dutch Association of Social Psychologists on 15 December 2017. All research in this chapter was presented in poster format at the Annual ASPO Conference of the Dutch Association of Social Psychologists on 14 December 2018.

ABSTRACT

Extant research suggests that non-visual information about another person can bias the way one visualizes that person's face in one's mind. How likely is this really to occur? In a series of four experiments, we tested under which circumstances non-visual behavioral information influences the visual mental representation of that person's face. We exposed participants to behavioral information about an unfamiliar person describing him as untrustworthy or trustworthy. We then visualized participants' mental representations of that person's face using reverse correlation. Behavioral information influenced mental representations of the face in line with this information in Study 3.1 (N=40), showing a strongly biased expected facial appearance. The effect disappeared in Study 3.2 (N=170) when participants were exposed to the person's face after the behavioral information. In Study 3.3 (N=200), we reduced exposure duration to the face and found mixed evidence. In Study 3.4 (N=200), we instructed participants to mentally visualize the expected facial appearance before brief exposure to the face. With this addition, behavioral information did influence mental representations. We conclude that effects of non-visual behavioral information on visualized mental representations of a seen face tend to be subtle if they arise at all. Moreover, we conclude that mentally visualizing the expected facial appearance combined with weakened sensory input from the actual face seems necessary to make such effects arise.

Keywords: social face perception, behavioral information, face prior, mental representation, reverse correlation, Bayesian models of perception

Facial appearance influences people's impressions of individuals (Antonakis & Eubanks, 2017; Klapper, Dotsch, van Rooij, & Wigboldus, 2016; Rule & Sutherland, 2017; Todorov, 2017; Todorov et al., 2015; Zebrowitz, 2017), leading to significant social consequences, such as court decisions (Blair et al., 2004; Porter et al., 2010; Wilson & Rule, 2015), financial investments (Chang et al., 2010; Rezlescu et al., 2012), electoral success (Antonakis & Eubanks, 2017; Olivola & Todorov, 2010a), personnel selection (Hassin & Trope, 2000), and more (see Todorov et al., 2015 for an overview). This happens even though face based impressions often lack accuracy (Antonakis & Eubanks, 2017; Efferson & Vogt, 2013; Todorov, 2017; Todorov et al., 2015). Moreover, people are influenced by facial appearances, even when more relevant information for person impressions is available (Chang et al., 2010; Olivola & Todorov, 2010b; Rezlescu et al., 2012) or when explicitly asked to ignore the face (Hassin & Trope, 2000).

A person's face can thus influence one's impressions of, and interactions with, that person. In the current contribution we investigate to what extent the reverse may also be true. That is, when we know about a person's behavior, does this influence how we mentally construct said person's face? People are known to actively and subjectively construct their perceptions of social targets (Bruner, 1957; Zaki, 2013). It is conceivable that they sometimes represent a person's face somewhat biased in their mind. Importantly however, the way a face is mentally represented can have consequences, such as how much money one entrusts to that person or whether one selects the person for interrogation or integration support (Kunst et al., 2017; Ratner et al., 2014). This suggests that, irrespective of what someone's face actually looks like, it is what that person's face looks like in one's mind that partly determines one's attitudes and behavior towards that person. Consequentially, it is relevant to understand the determinants of mental representations of faces.

In many cases, people have information about others before seeing their face (e.g. through gossip, information online, application procedures, perceiving someone from the back or a distance). Earlier work demonstrated that non-visual information or beliefs about people can influence visual mental representations of their faces (Brown-Iannuzzi, Dotsch, Cooley, & Payne, 2016; Dotsch, Wigboldus, & Van Knippenberg, 2013; Kunst, Dovidio, & Dotsch, 2017; Ratner, Dotsch, Wigboldus, van Knippenberg, & Amodio, 2014). These

studies focused on (1) beliefs about group members, not individuals, and (2) mental representations of faces participants had never seen, rather than faces participants had actually been exposed to. Do the same biasing effects still occur when dealing with an individual one has actually seen? It is conceivable that these effects disappear once one has seen the person's actual face, rendering the face in one's mind highly similar to the actual face. The biases would then only matter as long as one has not seen the person. Hassin and Trope (2000) suggest however that these biasing effects are incredibly persistent by showing that information about someone's personality changed one's perception of his facial features even when seeing his face.

The literature above suggests that it is possible that non-visual information about a person could influence one's visual mental representation of that person's face, even after seeing the face. Besides visual input from the actual face, non-visual information about the person could thus be a determinant of one's mental representation of that person's face. Extant research seems to focus on showing that this effect exists, but how likely is it really to happen? Moreover, Hassin and Trope (2000) focused only on a limited set of facial features a priori selected by the researchers, risking that they miss out on other, perhaps more relevant, facial features. In the present research, we investigated whether non-visual information about an *individual's* behavior, obtained *before* seeing the face, can influence the visual mental representation of that individual's face, even *after* exposure to the face. Using a data-driven methodology, we investigated the effect without restricting which facial features may be affected by the behavioral information. Importantly, besides investigating whether the effect occurs, we aim to shed more light on the *circumstances* (or moderating variables) under which the effect is more or less likely to occur. We approach these research questions with a theoretical view inspired by Bayesian models.

A Bayesian Inspired View on Social Face Perception

Social stimuli are complex. Their perception usually demands the integration of multiple cues from different modalities, which can be described by Bayesian inferential models (Zaki, 2013). Bayesian models have already been widely used to describe and predict object perception (Clark, 2013; Kersten & Yuille, 2003; Mamassian et al., 2002) and have recently won favor in informing theories about brain functioning, such as predictive coding (Clark, 2013; Friston, 2010; Kilner et

al., 2007). Bayesian models assume that people try to infer the most likely cause of their sensory input, using their prior knowledge and beliefs. This process is described using so-called prior, likelihood, and posterior distributions. To explain this process in the context of our research question, imagine a person called Alex meeting someone for the first time. One of the face evaluations made in a split second upon encountering someone, is how trustworthy the person appears (Marzi, Righi, Ottonello, Cincotta, & Viggiano, 2014; Todorov, Pakrashi, & Oosterhof, 2009; Willis & Todorov, 2006). In this example, let us take the trustworthiness appearance of the face as our inference of interest.

The so-called *prior distribution* depicts the probability of different scenarios, in this case a scale of facial appearance going from untrustworthy to trustworthy looking. Based on life experience, Alex may have learned that most people smile friendly when meeting someone and so expects to see a trustworthy looking face. The probabilities would thus show a peak somewhere on the trustworthy side of the prior distribution. However, imagine Alex heard that this particular person threatened another person and therefore expects this particular person to appear rather untrustworthy. Now the prior distribution is actually higher on the untrustworthy side. Then Alex meets the person. The *likelihood distribution* depicts the probability of one's sensory input, given the various facial trustworthiness appearances. If Alex' sensory input matches that of a(n) (un)trustworthy appearing face, the likelihood distribution will peak on the part of the scale matching this (un)trustworthy facial appearance. The two distributions are combined into a *posterior distribution*, which depicts the updated probability of the facial appearances. Using a decision rule (e.g. take the maximum of the distribution), a specific posterior belief can be generated about which facial trustworthiness appearance is the most likely cause of the sensory input (Mamassian et al., 2002).

When there is a mismatch between the prior and likelihood (e.g. Alex expected an untrustworthy looking face but sensory input matched that of a trustworthy looking face), the precision of distributions (the inverse of the variance) determines their relative influence on the posterior distribution (Edwards et al., 2012; Mamassian et al., 2002). If the sensory input from the trustworthy looking face is very clear, the peak of the likelihood distribution will be high

and small (i.e. precise), creating a large influence on the posterior distribution, and thus on Alex' experienced perception of the face.

Note that probabilities in Bayesian models are conditional (Zaki, 2013). If Alex believes people who threaten others always look untrustworthy, the expectation (i.e. prior probability) that this person will look untrustworthy is high for Alex. However, if Alex questions the reliability of the information about the person having threatened someone, or if Alex' association between threatening someone and looking untrustworthy is weak, the prior will be less precise. Likewise, if the reliability of Alex' sensory input is low (e.g. because the person was wearing sunglasses and/or a face mask, or Alex saw the person from a distance or in a poorly lit environment), the likelihood distribution will be less precise as well. The probabilities thus depend on how reliable each cue is considered in the specific context.

Based on such Bayesian models, we do not view the effect under investigation as something that simply does or does not happen. Instead, we view it as something that may happen, depending on the circumstances of the situation. In the present research, we aim to shed more light on these circumstances, or moderating variables. Based on our Bayesian inspired view, we hypothesize that earlier information about a person's behavior influences the mental representation of that person's seen face, if (1) the behavioral information indeed generates an expectation about that person's facial appearance (*face prior*), and (2) the face prior is relatively precise compared to the sensory input of the actual face. The first prediction is called the Face Prior Hypothesis and describes a prerequisite for an effect of behavioral information on mental representations to occur. The second prediction, called the Prior-Likelihood Balance Hypothesis, subsequently specifies to what extent we expect an effect to occur. In this view, the relative strength (or precision) of the face prior (expected facial appearance) and likelihood (based on sensory input from the actual face) determine to what extent an effect occurs.

If an effect is expected, Bayesian models could still predict the effect to be in one of two directions. Taken together, there are thus three possible ways how behavioral information may (or may not) influence mental representations of a person's seen face, which we will elaborate on now.

No effect: Sensory input overrules face priors

If the sensory input from the actual face is relatively strong (i.e. the likelihood distribution has high precision), face priors based on behavioral information should have no or minimal impact, and the mentally represented face should resemble the actual face accurately. This is in line with studies on the integration of inconsistent multisensory cues, which show that the most reliable cue is assigned most weight in perceptual judgment (Ernst & Banks, 2002; Fetsch, Pouget, DeAngelis, & Angelaki, 2012). Likewise, people update their impressions of others based on new information when this information is perceived to be diagnostic and reliable (Brannon & Gawronski, 2017; Cone & Ferguson, 2015; Cone, Mann, & Ferguson, 2017; Lammers, Gast, Unkelbach, & Galinsky, 2017). Even if behavioral information is considered a reliable cue for facial appearance (e.g. faces of people who tend to threaten / support others likely look more negative / positive), sensory input from the actual face should be a much more reliable cue for what the face out there looks like (e.g. if one's sensory input indicates that the face looks positive / negative, it is highly likely that the face indeed looks positive / negative). This suggests that the sensory input, being the most diagnostic and reliable cue concerning what the actual face looks like, should update any face priors one had beforehand. That is, of course, if the sensory input is indeed clear, resulting in a likelihood distribution with high precision.

Assimilation effect: A bias in the direction of the face prior

If face priors based on behavioral information are relatively strong (i.e. the prior distribution has high precision), they could bias one's mental representation of the face in the direction of the face prior (an assimilation effect). This is supported by research from Hassin and Trope (2000) in which personality information influenced ratings of a person's facial features, such as the fullness of the face and the shape of the chin. Dotsch, Wigboldus, and Van Knippenberg (2013) showed a similar assimilation effect of behavioral information about group members on the expected facial appearance of other members of that group. Note though that they investigated mental representations of the faces of unseen group members, not of the group members whose faces participants had already seen.

Other studies in social psychology demonstrated assimilation effects as well. For example, beliefs (in the form of stereotypes) about members from different social categories, like race or gender, lead participants to more readily recognize congruent than incongruent emotional expressions on members' faces (Bijlstra et al., 2014, 2010; Hugenberg, 2005). Moreover, when learning to couple behaviors to particular faces, face evaluations become affectively biased in line with the associated behavior (Bliss-Moreau, Barrett, & Wright, 2008; Falvello, Vinson, Ferrari, & Todorov, 2015).

Studies on perception have also shown assimilation effects. For instance, participants perceived the strength of bodily sensations consistent with their expectations and even perceived expected sounds that were not actually there (Edwards et al., 2012; Powers, Mathys, & Corlett, 2017). Finally, the selective accessibility model of Mussweiler (2003) states that, in most comparison situations, one likely focuses on similarities. If we interpret the face prior as a standard to which the actual face is compared, a focus on similarities leads to an assimilation effect, making the face appear more similar to the expected facial appearance than it actually is.

Contrast effect: A bias away from the face prior

If face priors based on behavioral information are relatively strong, it is also possible that they bias one's mental representation in the opposite direction of the face prior (a contrast effect). Although standard Bayesian models (e.g. Edwards et al., 2012; Kersten & Yuille, 2003) normally describe assimilation effects, adapted Bayesian models may predict contrast effects (Snyder, Schwiedrzik, Vitela, & Melloni, 2015). Moreover, the selective accessibility model (Mussweiler, 2003) predicts a contrast effect if the expected and actual facial appearance are extremely different from each other. In that case, Mussweiler (2003) states that in a first overall comparison, especially the differences between the two stand out, making the face appear more different from the expected facial appearance than it actually is. This is in line with aftereffects in face perception (Rhodes, 2017; Rhodes & Jeffery, 2006; Wincenciak, Dzhelyova, Perrett, & Barraclough, 2013), which may be applicable to the current research question if we assume that the behavioral information indeed generates a visual face prior, serving as the standard to which the sensory input from the actual face is compared.

The Present Research

With the present research, we aim to increase insight into whether non-visual behavioral information influences the visual mental representation of a seen face, and if so, under which circumstances and to what extent this influence can be expected to occur. In most if not all social situations, one of the most consequential person impressions is that of evaluating someone positively or negatively. The facial appearance dimension that comes closest to mere valence is trustworthiness, being one of the major underlying dimensions of person impressions (McAleer, Todorov, & Belin, 2014; Oosterhof & Todorov, 2008; Sutherland et al., 2013; Todorov et al., 2008), with significant social consequences (Chang et al., 2010; Porter et al., 2010; Rezlescu et al., 2012; Wilson & Rule, 2015). Therefore, we focused on perceived trustworthiness in our studies. We manipulated information about a male target person's behavior to depict him as trustworthy or untrustworthy and checked participants' trustworthiness evaluations of him in a manipulation check. Moreover, we measured the trustworthiness appearance of participants' mental representations of his face.[6]

We visualized mental representations of the face using a reverse correlation (RC) task (Brinkman, Todorov, & Dotsch, 2017; Dotsch & Todorov, 2012; Jack & Schyns, 2017). RC is a data-driven technique to create visual proxies of participants' mental representations without relying on participants' introspective abilities or on researchers' ideas about relevant facial features. It yields a classification image (CI), which is interpreted as an approximation of the participant's mental representation. Individual CIs (one for each participant) can be combined into group CIs (one for each experimental group). We report ratings of both individual and group CIs. We describe the technique in more detail in the Methods section.

We hypothesized that verbal information about a person's behavior affects the mentally represented facial trustworthiness of that person's seen face if (1) the behavioral information indeed generates a prior about the facial trustworthiness appearance of the person (face prior), and (2) this face prior is sufficiently strong

6 Note that we do not claim that these trustworthiness evaluations differ from mere valence evaluations. If preferred, one could read positively/negatively in place of trustworthy/untrustworthy. The point is whether and to what extent non-visual behavioral information influences visual mental representations of seen faces.

compared to the sensory input from the actual facial appearance. If a bias indeed occurs, we also aimed to discover whether the mental representation of the face is pulled towards the face prior (assimilation) or pushed away from the face prior (contrast). Given the mixed literature, we formulated no hypothesis regarding the direction of this effect.

Study 3.1 tested the first prediction that behavioral information can lead to a prior about facial trustworthiness appearance (Face Prior Hypothesis). Studies 3.2, 3.3, and 3.4 tested the second prediction (Prior-Likelihood Balance Hypothesis), by manipulating the relative strength of the face prior and the sensory input from the actual face. All studies were preregistered at the Open Science Framework[7] and received ethics approval from the institutional ethics committee. We report all manipulations, measures, and exclusions in the studies.

STUDY 3.1: DOES BEHAVIORAL INFORMATION CREATE A PRIOR ABOUT FACIAL APPEARANCE?[8]

We hypothesized that behavioral information would influence the expected facial appearance of the person (i.e. the face prior) such that participants would expect a more trustworthy looking face if the behavioral information depicted the person as trustworthy opposed to untrustworthy.

Method

Design and sampling plan

We manipulated behavioral information about a target person (untrustworthy / trustworthy; between-participants). The dependent variable was the degree of facial trustworthiness present in the CI as rated by an independent group of raters. We collected data until the Bayes factor (BF) for the behavioral information effect on facial trustworthiness was at least 10 ($BF_{10} \geqslant 10$ or BF_{01}

7 The preregistrations on the Open Science Framework can be found through the following links: Study 3.1 (https://osf.io/37s48), Study 3.2 (https://osf.io/6rb8m), Study 3.3 (https://osf.io/89gzs), and Study 3.4 (https://osf.io/6e5cm).

8 Study 3.1 was conducted after Study 3.3 (so the chronological order of studies was 3.2 – 3.3 – 3.1 – 3.4) but is presented first in this paper because it answers a different question than the other three studies.

⩾ 10),[9] with minimally 40 participants and maximally 200 participants after exclusion criteria were applied.

Because it would be costly to collect new ratings from a group of raters each time data should be analyzed to compute an intermediate BF, we quantified facial trustworthiness in CIs using a criterion CI of facial trustworthiness. The projection of any participant's CI on the criterion CI produces a trustworthiness score for that CI. These trustworthiness scores served as an indication of trustworthiness ratings as human raters would provide (see Chapter 2 for documentation and results on this criterion CI). We used the CI trustworthiness scores derived from this criterion CI to apply the stopping rule described above. Once we stopped data collection, we collected the ratings from human raters, which served as the dependent variable in our analyses.

Participants

Data in all studies were collected on Prolific Academic (https://www.prolific.ac) from Caucasian adults with self-reported normal or corrected-to-normal vision, with no restriction on country or education (see Table S3.1 and S3.2 for an overview of participants' residential countries and education levels for all studies in this chapter).[10] Out of 54 participants in this study, 14 participants were excluded based on preregistered criteria aimed at removing unmotivated participants: 5 participants failed the name-behaviors association check (see below), 5 participants failed the attention check (see below), and 4 participants had a median reaction time in the RC task below 400 milliseconds.[11] The

9 BFs indicate that the data provide more evidence for the null model or for the alternative model, or that they do not discriminate the models (Dienes, 2016). A BF of 10 is considered "strong" evidence for one model over the other (Jeffreys, 1961).

10 We aimed to increase generalizability of findings by including participants from diverse geographical, educational, and age backgrounds. We did, however, select only Caucasian participants in each study to investigate the effect of interest without potential intergroup biases (as the base face and target face used in our studies were Caucasian). Although we see no reason to assume that the influence of behavioral information on mentally represented faces works differently for non-Caucasian people, we cannot exclude that possibility based on our sample.

11 In contrast to 500 milliseconds in the other studies, we selected 400 milliseconds as exclusion criterion in the current study. Our reasoning was that participants in the current study would probably base their face choices on the mere question of which face looks more (un)trustworthy, whereas participants in the other studies should be choosing the face that looks more like the target person's actual face. We expect the first choice to be easier and therefore quicker made than the latter choice. Hence the difference of 100 milliseconds in our exclusion criterion.

final sample consisted of 40 participants (24 women, 16 men, M_{age} = 36.30, SD_{age} = 11.05).

Procedure and materials

The study was designed and hosted online using Gorilla (https://gorilla.sc/). Participants were informed that the study may terminate after a few tasks and that they would be paid for participation up until that point if that happened. After providing informed consent, participants were asked to ensure a quiet environment without distractions before moving on to the following tasks in chronological order.

Behavioral information

Participants read behavioral information about 4 different men (A. Brown, L. Harris, H. Young, and F. Taylor). They were instructed to form an impression about each person. The order of names was random, except that F. Taylor, the target person, was always last. Each person was described with 10 behavioral descriptions, presented one by one in random order in the middle of the screen below the person's name. The behavioral descriptions were validated in a pilot study (see Appendix 3A). In the *untrustworthy* condition, F. Taylor was paired with the behavioral descriptions from the untrustworthy set. In the *trustworthy* condition, F. Taylor was paired with the trustworthy set. The task was self-paced with at least 1 second per behavioral description.

Name-behaviors association check

To check whether participants remembered which behaviors were paired with F. Taylor, they selected the set of behaviors they believed were performed by F. Taylor. The 4 sets were presented from left to right in a different order than in the previous task. If participants selected an incorrect set, they were excluded from further participation and were paid.

Attention check

On one instruction page, as attention check, we included an Instructional Manipulation Check (Oppenheimer, Meyvis, & Davidenko, 2009). Participants were instructed to click the title of the instructions page or press the *A* key to proceed instead of clicking the continue button. If participants nevertheless

clicked the continue button, they were excluded from further participation and were paid.

Reverse correlation task

To visualize their mental representations of F. Taylor, participants completed a two-images forced choice RC task. In this task, participants selected the face (out of two faces) they would say most likely belonged to F. Taylor, on 500 trials. Participants could take a break after every 100 trials if they wanted to. All stimuli consisted of the same base image, which was the grayscale average of the average male and female faces of the Averaged Karolinska Directed Emotional Faces database (Lundqvist & Litton, 1998), with random noise superimposed on the image (Figure 3.1). The noise was unique for each stimulus and generated following the procedure described in Dotsch and Todorov (2012). On each trial, the noise pattern for the left stimulus was the opposite of the noise pattern for the right stimulus. Stimuli were generated in R (R Development Core Team, 2016) using version 0.3.4.1 of the rcicr package (Dotsch, 2016).

a **b** **c**

Figure 3.1. Base face used in the reverse correlation task (a), with a random noise pattern (b) and its inverse (c) superimposed on the base face. The base face is the average of images FNES and MNES of the Averaged Karolinska Directed Emotional Faces database (Lundqvist & Litton, 1998).

Manipulation check

Participants evaluated F. Taylor on several scales, amongst which the trustworthiness scale, ranging from -4 (*untrustworthy*) to 4 (*trustworthy*; see Appendix 3B for all scales). The other scales were added to disguise our interest in trustworthiness.

Questionnaire

Participants answered questions about their experience during the experiment (see Appendix S-3A for all questions), demographics (gender, age, native language, country of residence, and highest completed education), and proficiency in English. Participants could also leave remarks.

Last, participants were redirected to Prolific and paid.

CI generation and trustworthiness evaluation

CI generation

We generated CIs using version 0.4.0 of the rcicr package (Dotsch, 2017). Per participant, all selected noise patterns were averaged to generate individual classification patterns (CPs) of F. Taylor. These were visualized as faces after being scaled and superimposed on the base image, resulting in the individual CIs. Group CPs were generated per behavioral information condition by averaging the raw individual CPs per condition. Like the individual CPs, these group CPs were visualized into group CIs (Figure 3.2).

Individual CIs rating task

Twenty-one Caucasian raters (9 women, 11 men, 1 androgynous, M_{age} = 33.19, SD_{age} = 10.45) evaluated the 40 individual CIs, in random order, on a 9-point scale ranging from -4 (*very untrustworthy*) to 4 (*very trustworthy*). Raters evaluated a subset of 20 CIs again, without knowing they had already rated these images. The intraclass correlation (ICC3,1 ; Shrout & Fleiss, 1979) between a rater's initial and repeated judgment of these CIs served as indication of a rater's consistency, which varied highly (M = .53, SD = .23).[12] We quantified the amount of trustworthiness signal present in individual CIs using the initial ratings. First, initial ratings were standardized per rater to make them comparable between raters.[13] Subsequently, the average trustworthiness rating was computed for each individual CI.

12 We preregistered to compute correlations. We specifically computed intraclass correlations Model 3, Form 1, because it is considered a fitting indicator of intrarater consistency (Shrout & Fleiss, 1979; Trevethan, 2017).

13 For each study, using standardized or unstandardized ratings did not change the conclusions.

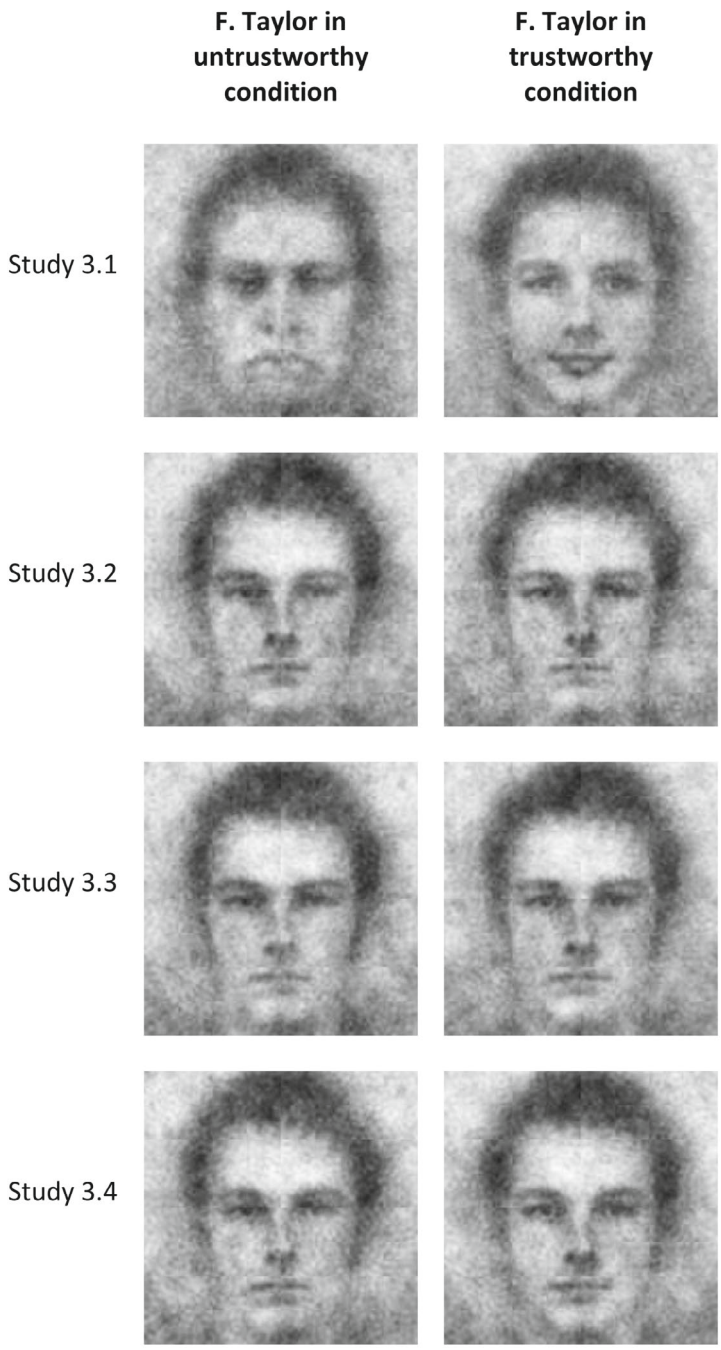


Figure 3.2. Visualization of the group mental representations of F. Taylor's facial appearance when participants had not seen (Study 3.1) or had seen (Study 3.2, 3.3 and 3.4) F. Taylor's face and participants had read the descriptions depicting F. Taylor as untrustworthy (left column) or trustworthy (right column).

Results

Confirmatory analyses

In line with our preregistration, all Bayesian analyses in this paper were conducted using the default prior settings in JASP (JASP Team, 2020). For these analyses, we report 95% credible intervals of the estimated effect sizes between brackets. We encourage readers to interpret the BF as a continuous measure of evidence for one model over another model. In Bayesian ANOVA's, reported Inclusion BFs ($BF_{Inclusion}$) compare all models that contain the effect of interest (e.g. an interaction effect) with all equivalent models stripped of that effect. For Study 3.1, we performed one-sided Bayesian independent samples *t*-tests using the default Cauchy prior settings in JASP with $r = 1/\sqrt{2}$, truncated such that only negative effect size values were allowed. A negative effect size indicates that the untrustworthy condition scores on average lower than the trustworthy condition.

Manipulation check

As intended, F. Taylor was evaluated less trustworthy in the untrustworthy (M = -2.95, SD = 1.54) than trustworthy condition (M = 3.35, SD = 1.46, posterior median effect size δ = -4.06, 95% credible interval [-5.25; -2.94], BF_{-0} = 9.23 x 10^{12}). Because the assumption of normality was violated, we tested the robustness of the result with a Bayesian Mann-Whitney U test, which also provided more evidence for the alternative hypothesis that F. Taylor was evaluated less trustworthy in the untrustworthy condition (BF_{-0} = 1025.64).

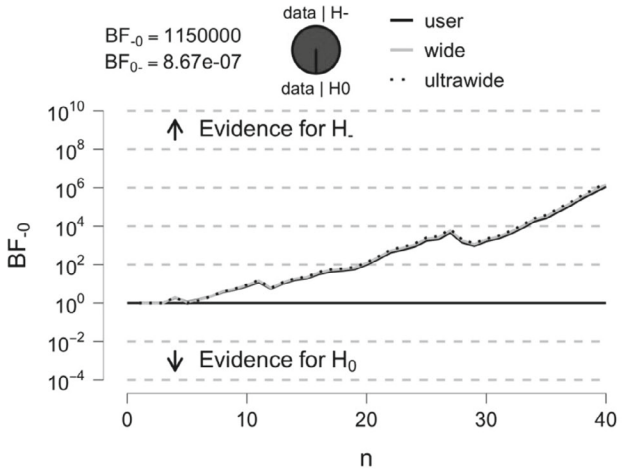
Individual CI trustworthiness

In line with the Face Prior Hypothesis, trustworthiness ratings were lower for CIs of F. Taylor in the untrustworthy (M = -.55, SD = .34) than in the trustworthy condition (M = .55, SD = .35, δ = -3.05 [-4.01; -2.12], BF_{-0} = 3.34 x 10^{9}).[14] Figure 3.3 shows how the BF developed as sample size increased both for the projection-based trustworthiness scores, which were used as data collection stopping rule,

14 Two participants may have misremembered the behaviors linked with F. Taylor later in Study 3.1. Their verbal descriptions of their expectations of him were opposite in valence to the behavioral information presented to them. Excluding these participants did not change the conclusions (BF_{-0} = 4.05 x 10^{10}).

and for the human trustworthiness ratings, which were used as dependent variable in the current analysis.

(A) Projection-based trustworthiness scores



(B) Human trustworthiness ratings

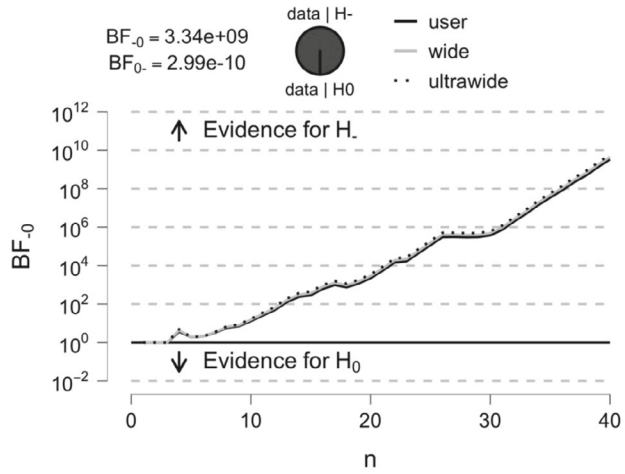


Figure 3.3. Sequential analyses tracking the Bayes factor (BF) as sample size increases for a variety of prior widths (*r*: as used in the analysis / wide / ultrawide) for (A) the projection-based trustworthiness scores, used as data collection stopping rule, and (B) the human trustworthiness ratings, used as the dependent variable of interest in Study 3.1. The user prior is the JASP default prior.

Exploratory analyses

Appendix 3C presents a mediation analysis investigating whether the effect of behavioral information on the expected facial trustworthiness appearance may be fully mediated by the trustworthiness impression of F. Taylor.

The results provide evidence for the idea that valenced behavioral information can generate priors (expectations) about facial trustworthiness appearance. It seems that the non-visual behavioral information led to activation of specific visual facial characteristics. These were reflected in participants' mental representations of F. Taylor's face, which looked more trustworthy in the trustworthy opposed to untrustworthy condition.

STUDY 3.2: DOES BEHAVIORAL INFORMATION INFLUENCE MENTAL REPRESENTATIONS OF SEEN FACES?

In Study 3.2, we tested whether behavioral information influences participants' mental representations of F. Taylor's face when they have seen his actual face. We hypothesized that it would. Given that the literature supports predictions for assimilation (e.g. Hassin & Trope, 2000) as well as contrast effects (e.g. Rhodes & Jeffery, 2006) when the actual face is involved, we had no expectation about the direction of the effect.

Method

The method was identical to Study 3.1, except where indicated below.

Design and sampling plan

There was a second independent variable: manipulation check (present / absent; between-participants). The data collection stopping rule was applied before exclusion criteria were applied.

Participants

Out of 249 participants[15], 59 participants were excluded based on preregistered criteria: 31 participants failed the name-behaviors association check, 22 failed the attention check, and 6 had a median reaction time in the RC task below 500 milliseconds. Moreover, 19 participants were excluded because of a server error during the experiment. One participant did not complete the experiment. The final sample consisted of 170 participants (65 women, 105 men, M_{age} = 32.32, SD_{age} = 10.00).

15 After excluding participants based on our preregistered criteria, the four conditions had very unequal cell sizes. We collected additional data to get approximately equal cell sizes, ending at 249 instead of 200 participants.

Procedure and materials

Participants performed the behavioral information task, manipulation check, name-behaviors association check, F. Taylor's face information, attention check, RC task, and questionnaire, in this order.

Manipulation check

The manipulation check, which measures participants' impression of F. Taylor's trustworthiness, was administered right after the behavioral information. Because the manipulation check itself may modulate the effect of behavioral information on the mentally represented face, only half of the participants received it, orthogonal to behavioral information condition. These participants evaluated the four men in the same order as they were presented during the behavioral information phase.

F. Taylor's face

Participants viewed a picture of F. Taylor's face (Figure 3.4), which was the grayscale version of a male face from the Radboud Faces Database (Langner et al., 2010). The face has a neutral expression and scored on average -.05 (*SD* = 1.05) on a 5-point trustworthiness scale ranging from -2 (*very untrustworthy*) to 2 (*very trustworthy*). F. Taylor's face was presented for 10 s, which is twice the presentation time sufficient for relatively high rates of face recognition (Bower & Karlin, 1974). Participants were instructed about this time limit beforehand, were told to take a good look at the face, and were aware that they subsequently needed to select the face that looked most like F. Taylor in 500 sets of faces.

Figure 3.4. Face presented as F. Taylor's face. The face is taken from the Radboud Faces Database (Langner et al., 2010).

RC task

We slightly altered the instruction to acknowledge that participants had seen F. Taylor's face. Instead of asking "Which face would you say most likely belongs to F. Taylor?" participants were asked "Which face looks most like F. Taylor?".

CI generation and trustworthiness evaluation

Individual CI rating task

Twenty Caucasian raters (9 women, 11 men, M_{age} = 30.30, SD_{age} = 10.93) evaluated all 170 individual CIs, and a subset of 30 CIs again (rater consistency: M = .32, SD = .23). The lower average consistency than in the previous study indicates that it was harder for raters to judge the trustworthiness appearance of the faces in the individual CIs in this study.

Results

Confirmatory analyses

Manipulation check

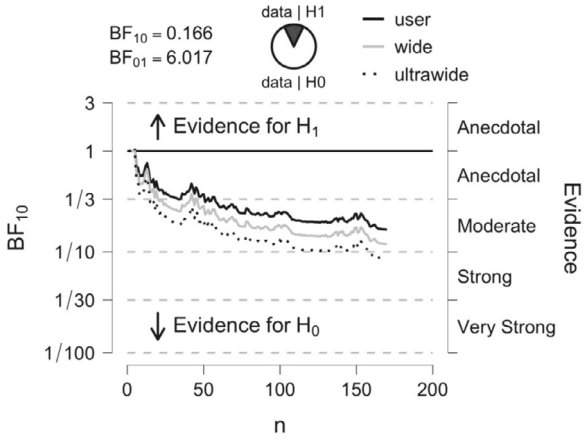
We conducted the same *t*-test as in Study 3.1. F. Taylor was again evaluated as less trustworthy in the untrustworthy (M = -3.57, SD = .89) than trustworthy condition (M = 3.61, SD = .78, δ = -8.49 [-9.91; -7.16], BF_{-0} = 5.20 x 10^{52}). The assumption of normality was again violated. Here too, a Bayesian Mann-Whitney U test provided more evidence for the alternative hypothesis (BF_{-0} = 568508.01).

Individual CI trustworthiness

CI trustworthiness was predicted by behavioral information (untrustworthy/ trustworthy; between-participants) and manipulation check (present / absent; between-participants) in a Bayesian ANOVA. Because we had no specific expectations about sizes of effects, we used the default JASP prior settings (*r* scale fixed effects = 0.5; *r* scale random effects = 1). Contrary to our hypothesis, the data were 5.95 times more likely under the null model than under the model including the effect of behavioral information (BF_{01} = 5.95), implying that trustworthiness ratings did not differ between CIs of F. Taylor in the behavioral information conditions. The effect also did not seem to emerge for only those participants who had received the manipulation check, as there

was no compelling interaction effect ($BF_{Inclusion}$ = .50[16]). Figure 3.5 shows how the BF developed as sample size increased both for the effect of behavioral information on the projection-based trustworthiness scores, which were used as data collection stopping rule, and on the human trustworthiness ratings, which were used as dependent variable in the current analyses.

(A) Projection-based trustworthiness scores



(B) Human trustworthiness ratings

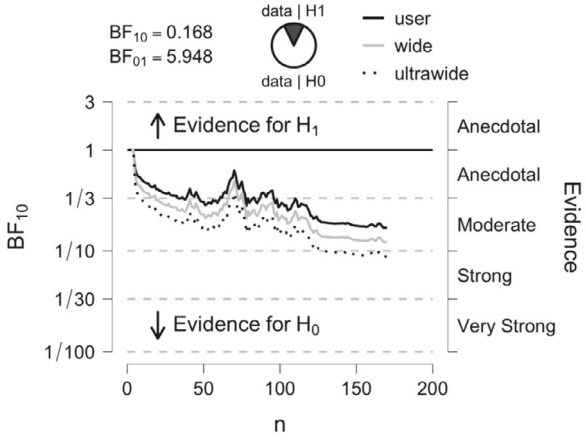


Figure 3.5. Sequential analyses tracking the Bayes factor (BF) as sample size increases for a variety of prior widths (*r*: as used in the analysis / wide / ultrawide) for the effect of behavioral information on (A) the projection-based trustworthiness scores, used as data collection stopping rule, and (B) the human trustworthiness ratings, used as the dependent variable of interest in Study 3.2. The user prior is the JASP default prior.

16 Which is equal to 1 / .50 = 2.00 against including the interaction effect.

Exploratory analyses

Group CI trustworthiness

As described in Chapter 2, individual CI ratings in this study were influenced by CI picture quality. The noisier the CI, the less trustworthy raters evaluated the CI. Because group CIs have better signal-to-noise ratio than individual CIs, we decided to collect trustworthiness ratings of group CIs as well in the upcoming studies. See Study 3.3 and 3.4 for exploratory and confirmatory analyses on trustworthiness ratings of Study 3.2's group CIs, showing that the group CI results are in line with the individual CI results for Study 3.2.

Contrary to our expectations, the data of Study 3.2 provided more evidence for the null than alternative hypothesis. Behavioral information did not bias participants' mental representations of F. Taylor's face when his face was presented for 10 s.

STUDY 3.3: DOES BEHAVIORAL INFORMATION INFLUENCE MENTAL REPRESENTATIONS OF BRIEFLY SEEN FACES?

Bayesian models suggest that the predicted effect of behavioral information on mentally represented faces should emerge when sensory input from the face is relatively weak compared to the face prior. Could it be that sensory input was relatively strong in Study 3.2? Participants had an unobstructed view of the face for 10 s and were instructed beforehand to take a good look at the face. Indeed, participants still remembered F. Taylor's facial characteristics in the questionnaire at the end of the experiment (e.g. "lips pointed down, big round ears, dark circle around eyes"). This suggests that sensory input from the actual face may have been rather strong in Study 3.2.

In Study 3.3, we aimed to investigate whether evidence may shift in favor of the alternative opposed to the null model when we manipulate sensory input from the face to be weaker. There are multiple ways conceivable in which this could be achieved. For Study 3.3, we chose to reduce the strength of the sensory input in two ways. First, we showed participants F. Taylor's face only briefly, namely 100 ms (vs. 10 s in Study 3.2). Second, instead of creating a goal to accurately remember the face, we simply informed participants that they would

catch a very brief glimpse of F. Taylor's face. We hypothesized that behavioral information would influence participants' mental representations of F. Taylor's face and had no expectation regarding the direction of the effect.

Method

The method was identical to Study 3.1, except where indicated below.

Participants

Out of 297 participants, 91 participants were excluded based on preregistered criteria: 35 participants failed the name-behaviors association check, 40 failed the attention check, and 16 had a median reaction time in the RC task below 500 milliseconds. Moreover, 6 participants did not complete the experiment, rendering their data incomplete. The final sample consisted of 200 participants (95 women,105 men, M_{age} = 35.90, SD_{age} = 12.11).

Procedure and materials

Participants performed the behavioral information task, name-behaviors association check, F. Taylor's face information, attention check, RC task, manipulation check, and questionnaire, in this order.

F. Taylor's face

F. Taylor's face, which was the same as in Study 3.2, was presented for 100 ms, which is long enough to form trustworthiness judgments (Todorov et al., 2009). Participants were informed beforehand that they would get a brief glance at the face. They were then presented with a fixation cross in the middle of the screen for 500 ms, followed by the face for 100 ms, and a mask for 500 ms.

RC task

Participants were asked to select the face that looked most like F. Taylor *to them* ("Which face looks most like F. Taylor to you?"). With this slight revision, we tried to subtly decrease the feeling that participants should remember all facial features accurately.

CI generation and trustworthiness evaluation

Individual CI rating task

Out of 22 raters, 2 raters were excluded based on the preregistered criterion that their median reaction times were below 500 milliseconds. The final sample consisted of 20 Caucasian raters (10 women, 10 men, M_{age} = 35.90, SD_{age} = 12.84), who evaluated all 200 individual CIs, and a subset of 20 CIs again (rater's consistency: M = .44, SD = .25).

Group CI forced choice and rating task

For explorative purposes, 49 Caucasian raters (26 women, 23 men, M_{age} = 29.78, SD_{age} = 9.36) evaluated the group CIs of both Study 3.2 and 3.3. First, they selected for each study which of the 2 group CIs they thought looked most trustworthy (the forced choice task). Second, they rated each group CI on a 9-point scale ranging from -4 (*very untrustworthy*) to 4 (*very trustworthy*).

Results

Confirmatory analyses

We conducted similar *t*-tests as in Study 3.1, except that the *t*-test for CI trustworthiness was two-sided. Consequentially, the default Cauchy prior settings in JASP for this test stated $r = 1/\sqrt{2}$ without truncation, allowing for both negative and positive effect size values.

Manipulation check

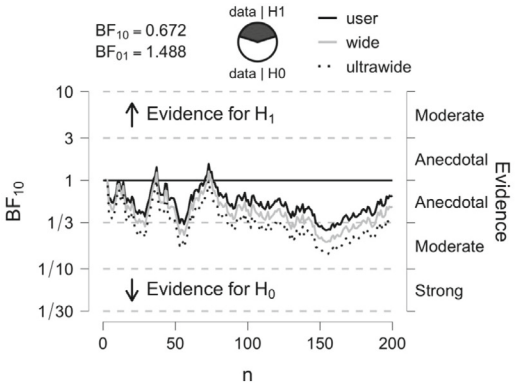
F. Taylor was again evaluated as less trustworthy in the untrustworthy (M = -2.99, SD = 1.65) than trustworthy condition (M = 2.91, SD = 1.77, δ = -3.43 [-3.87; -3.00], BF_{-0} = 1.19 x 10^{58}). Due to violation of the assumption of normality, a Bayesian Mann-Whitney U test was conducted to test the robustness of the results. Indeed, the data provided more evidence for the alternative hypothesis (BF_{-0} = 1.91 x 10^{8}).

Individual CI trustworthiness

Trustworthiness ratings provided moderate evidence against a difference between the untrustworthy (M = -.04, SD = .47) and trustworthy condition

(M = .04, SD = .50, δ = -.16 [-.43; .11], BF_{01} = 3.25).[17] Figure 3.6 shows how the BF developed as sample size increased both for the projection-based trustworthiness scores, which were used as data collection stopping rule, and for the human trustworthiness ratings, which were used as dependent variable in the current analyses.

(A) Projection-based trustworthiness scores



(B) Human trustworthiness ratings

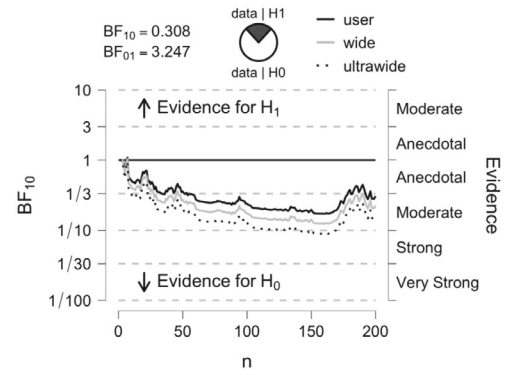


Figure 3.6. Sequential analyses tracking the Bayes factor (BF) as sample size increases for a variety of prior widths (*r*: as used in the analysis / wide / ultrawide) for (A) the projection-based trustworthiness scores, used as data collection stopping rule, and (B) the human trustworthiness ratings, used as the dependent variable of interest in Study 3.3. The user prior is the JASP default prior.

17 Twelve participants may have misremembered the behaviors linked with F. Taylor later in the experiment. Their ratings on the manipulation check were opposite in valence to the behavioral information presented to them. Excluding these participants slightly increased evidence for the null (BF_{01} = 4.03). Moreover, 10 participants (of which one participant was also in the list of participants who may have misremembered the behaviors) indicated they had not seen F. Taylor's face during the face presentation task. Removing these participants on top of those who may have misremembered the behaviors still resulted in moderate evidence against a difference between the untrustworthy (M = .01, SD = .46) and trustworthy condition (M = .05, SD = .50, δ = -.09 [-.37; .20], BF_{01} = 5.14).

Exploratory analyses

Group CI trustworthiness Study 3.2 and 3.3

For Study 3.2's group CIs, only 19 out of 49 raters selected the CI from the trustworthy condition as looking most trustworthy, BF_{10} Poisson = 2.16. For Study 3.3's group CIs, however, 39 out of 49 raters selected the CI from the trustworthy condition, BF_{10} Poisson = 4319.63, which is in line with an assimilation effect.

The ratings further substantiated these findings. We conducted a Bayesian repeated measures ANOVA predicting trustworthiness ratings by behavioral information condition of the group CI (untrustworthy / trustworthy; within-participants) and study (Study 3.2 / Study 3.3; within-participants), using the default prior settings. The data provided evidence for an interaction effect ($BF_{Inclusion}$ = 41561.79). Bayesian paired-samples *t*-tests provided evidence that Study 3.2's group CIs did not differ on trustworthiness ($M_{untrustworthy}$ = .39, $SD_{untrustworthy}$ = 1.61 vs. $M_{trustworthy}$ = .29, $SD_{trustworthy}$ = 1.53, δ = .05 [-.22; .32], BF_{01} = 6.00), whereas for Study 3.3, trustworthiness ratings were higher for the group CI from the trustworthy (M = 1.71, SD = 1.67) than untrustworthy condition (M = -.18, SD = 1.75, δ = -.94 [-1.28; -.60], BF_{10} = 904715.94). The data of Study 3.2 thus showed no effect, whereas the data of Study 3.3 showed an assimilation effect.[18]

Discussion

We hypothesized that a face presentation of 100 ms would render the sensory input from the face relatively weak, giving the face priors a chance to influence the mental representation. Although the data on the group CIs, which can be interpreted as the average mental representation of all participants in an

18 It would be interesting to see whether the effect on the group CIs may be driven by the few participants who had not seen F. Taylor's face during the face presentation task. Because the group CIs are the averages of the individual CIs we do not have trustworthiness ratings of group CIs without the individual CIs of these 10 participants. However, we generated the group CIs including and excluding the individual CIs of these 10 participants and inspected them visually. It appeared that excluding these individual CIs from the group CIs made no difference, suggesting that it would not change the conclusions for the group CI trustworthiness analyses (see Appendix S-3B). We did the same for Study 3.4, where 8 participants indicated they had not seen F. Taylor's face during the face presentation task. Again, excluding these participants did not seem to have an impact on the group CIs and consequently on the conclusions for the group CI trustworthiness analyses (see Appendix S-3B).

experimental group, provided extreme evidence for an assimilation effect, the data on the individual CIs, which represent mental representations of individual participants, did not.

This mixed evidence can result from two things. First, the effect was subtly present in most individual CIs. When averaging across participants, noise from different individual CIs canceled each other out, bringing out the effect more clearly in the group CIs. Second, the effect was strongly present in a few individual CIs that contained less noise than the other individual CIs, thereby largely determining the group CIs. Either way, the different signal between conditions must have been present in some (strongly) or most (subtly) individual CIs to materialize so clearly in the group CIs. This difference between trustworthiness conditions was absent in Study 3.2, where both individual and group CI ratings provided more evidence for the null hypothesis. Although the overall data of Study 3.3 shifted somewhat in the direction of an effect, they provided mixed, and therefore not yet convincing, support for an effect of behavioral information on mental representations.

STUDY 3.4: DOES BEHAVIORAL INFORMATION INFLUENCE MENTAL REPRESENTATIONS WHEN MENTALLY VISUALIZING THE FACE PRIOR BEFORE BRIEFLY SEEING THE FACE?

The slight shift in support in the direction of an effect from Study 3.2 to Study 3.3 carefully suggests that the relative strength of the face prior and sensory input of F. Taylor's face may indeed matter for the extent to which behavioral information might influence mental representations. Although this 'extent' may be substantial on the level of group mental representations, the individual CI data suggest that it can be considered non-existent on the level of individual mental representations. Moreover, the group CI analyses were exploratory in nature. In Study 3.4, we planned to conduct confirmatory analyses on the group CIs.

Moreover, we planned to further strengthen the face prior based on the behavioral information. Although Study 3.1 showed that behavioral information can generate face priors, participants may form stronger face priors when they are triggered to think about facial appearance (which they were during the

RC task in Study 3.1). If true, the effect of behavioral information on mental representations should become stronger when participants mentally visualize their expectation of the face before viewing F. Taylor's face. We tested this in Study 3.4. We hypothesized again to find an effect of behavioral information on mental representations.[19]

Method

The method was identical to Study 3.1, except where indicated below.

Participants

Out of 299 participants, 96 participants were excluded based on preregistered criteria: 36 participants failed the name-behaviors association check, 39 failed the attention check, 19 had a median reaction time in the RC task below 500 milliseconds, and 2 participants did not describe a face when describing their face prior (see below). Moreover, 3 participants did not complete the experiment, rendering their data incomplete. The final sample consisted of 200 participants (126 women, 74 men, M_{age} = 27.43, SD_{age} = 4.31).

Procedure and materials

These were identical to Study 3.3, except where indicated below.

Face prior visualization

After the behavioral information, but before F. Taylor's face information, participants were instructed to form an expectation of F. Taylor's face and visualize it in their mind as clearly as they could. To ensure that participants did this, they were asked to write down in their own words what they expected F. Taylor's face to look like.

19 The results of Study 3.3 suggest that we should expect an assimilation effect in Study 3.4. However, we deemed it possible that mentally visualizing the expected facial appearance more strongly would make the differences between the expected and actual face stand out more, which may result in a contrast effect. Therefore, we still did not formulate a hypothesis about the direction of the effect in Study 3.4.

CI generation and trustworthiness evaluation

Individual CI rating task

Twenty Caucasian raters (11 women, 9 men, M_{age} = 31.95, SD_{age} = 10.14) evaluated all 200 individual CIs, and a subset of 20 CIs again (rater's consistency: M = .35, SD = .23).

Group CI forced choice and rating task

Now for confirmative purposes, 41 Caucasian raters evaluated the group CIs of Study 3.2, 3.3, and 3.4. One participant self-declared to have insufficient English proficiency. The final sample consisted of 40 raters (26 women, 14 men, M_{age} = 34.70, SD_{age} = 11.17).

Results

Confirmatory analyses

The conducted *t*-tests were identical to Study 3.3.

Manipulation check

F. Taylor was again evaluated less trustworthy in the untrustworthy (M = -3.34, SD = 1.23) than trustworthy condition (M = 2.99, SD = 1.21, δ = -5.16 [-5.74; -4.57], BF_{-0} = 2.69 x 10^{86}). A Bayesian Mann-Whitney U test again supported this finding (BF_{-0} = 2.65 x 10^{8}).

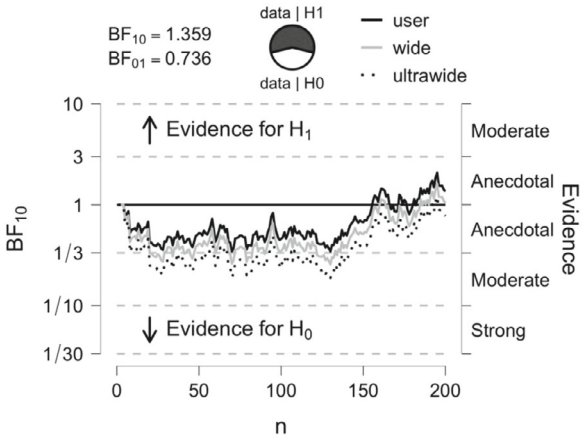
Individual CI trustworthiness

Trustworthiness ratings provided moderate evidence for CIs of F. Taylor appearing less trustworthy in the untrustworthy (M = -.09, SD = .43) than trustworthy condition (M = .09, SD = .44, δ = -.38 [-.66; -.11], BF_{10} = 6.52), which is in line with an assimilation effect.[20] Figure 3.7 shows how the BF developed as sample size increased both for the projection-based trustworthiness

20 Four participants may have misremembered the behaviors linked with F. Taylor later in Study 3.4. Their ratings on the manipulation check were opposite in valence to the behavioral information presented to them. Excluding these participants would strengthen the conclusions (BF_{10} = 12.74). However, 8 participants (of which one participant was also in the list of participants who may have misremembered the behaviors) indicated they had not seen F. Taylor's face during the face presentation task. Excluding these participants on top of the 4 who misremembered the behaviors, weakened the effect ($M_{untrustworthy}$ = -.07, $SD_{untrustworthy}$ = .42 vs. $M_{trustworthy}$ = .10, $SD_{trustworthy}$ = .43, δ = -.37 [-.65; -.08], BF_{10} = 4.40), resulting again in moderate evidence for an assimilation effect.

scores, which were used as data collection stopping rule, and for the human trustworthiness ratings, which were used as dependent variable in the current analyses.

(A) Projection-based trustworthiness scores



(B) Human trustworthiness ratings

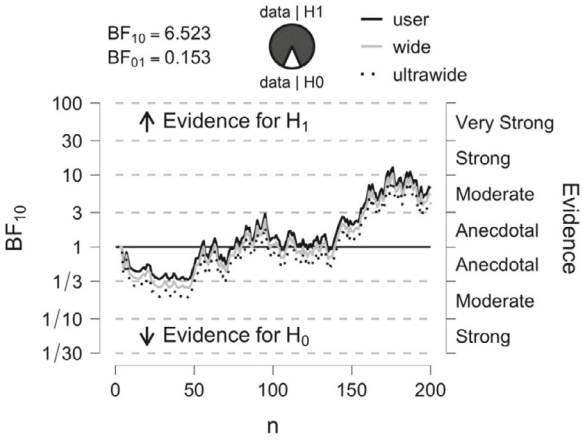


Figure 3.7. Sequential analyses tracking the Bayes factor (BF) as sample size increases for a variety of prior widths (*r*: as used in the analysis / wide / ultrawide) for (A) the projection-based trustworthiness scores, used as data collection stopping rule, and (B) the human trustworthiness ratings, used as the dependent variable of interest in Study 3.4. The user prior is the JASP default prior.

Group CI trustworthiness Study 3.2, 3.3, and 3.4

The forced choice data showed increasing consensus across studies on which group CI looked most trustworthy. For Study 3.2's group CIs, 18 out of 40 raters selected the CI from the trustworthy condition as looking most trustworthy, BF_{01} Independent multinomial = 2.19. For Study 3.3's group CIs, 32 out of 40 raters

selected the trustworthy condition CI, BF_{10} Independent multinomial = 527.34, and for Study 3.4's group CIs, 39 out of 40 raters selected the trustworthy condition CI as most trustworthy, BF_{10} Independent multinomial = 6.12 x 10^8.

The ratings confirmed this trend. We conducted a Bayesian repeated measures ANOVA predicting trustworthiness ratings by behavioral information condition of the group CI (untrustworthy / trustworthy; within-participants) and study (Study 3.2 / Study 3.3 / Study 3.4; within-participants), using the default prior settings. The data showed support for an interaction effect ($BF_{Inclusion}$ = 2.09 x 10^8). Bayesian paired-samples *t*-tests confirmed that Study 3.2's group CIs did not differ on trustworthiness ($M_{untrustworthy}$ = -.10, $SD_{untrustworthy}$ = 1.63 vs. $M_{trustworthy}$ = -.45, $SD_{trustworthy}$ = 1.68, δ = .17 [-.13; .47], BF_{01} = 3.10), whereas for Study 3.3, the trustworthy condition group CI of F. Taylor (M = 1.08, SD = 1.46) was rated as more trustworthy than the untrustworthy condition group CI (M = -.55, SD = 1.66, δ = -.75 [-1.11; -.40], BF_{10} = 1589.85). This replicates the findings of the group CI ratings collected after Study 3.3. The *t*-test for Study 3.4 showed evidence for an assimilation effect as well ($M_{trustworthy}$ = 2.10, $SD_{trustworthy}$ = 1.39 vs. $M_{untrustworthy}$ = -.70, $SD_{untrustworthy}$ = 1.76, δ = -1.30 [-1.74; -.87], BF_{10} = 5.43 x 10^7). A follow-up Bayesian repeated measures ANOVA on the data of only Study 3.3 and 3.4 provided evidence for an interaction effect ($BF_{Inclusion}$ = 6.08), suggesting that behavioral information had a larger effect in Study 3.4 than 3.3.

See Table 3.1 for an overview of all *t*-test results testing the effect of behavioral information on the mentally represented facial trustworthiness for Studies 3.1-3.4. In line with the Prior-Likelihood Balance Hypothesis, the effect sizes increase from Study 3.2 through 3.4.

Table 3.1. Bayesian t-test results of the impact of behavioral information (untrustworthy / trustworthy) on trustworthiness ratings of the mentally represented facial appearance of F. Taylor, as rated by independent groups of raters. Results are shown for all studies and for ratings of individual and group classification images (CIs). The alternative hypothesis states that the untrustworthy and trustworthy condition differ from each other on the trustworthiness ratings.

		Untrustworthy	Trustworthy			
Study	**CIs**	***M (SD)***	***M (SD)***	**BF_{10}**	**BF_{01}**	**Posterior effect size δ**
3.1	Individual	-0.55 (0.34)	0.55 (0.35)	1.67×10^{9}	5.99×10^{-10}	-3.05 [-4.01; -2.12]
	Individual[a]	-0.57 (0.34)	0.60 (0.29)	2.03×10^{10}	4.93×10^{-11}	-3.53 [-4.59; -2.48]
3.2	Individual	-0.01 (0.43)	0.01 (0.46)	0.17	5.95	-0.02 [-0.31; 0.27]
	Group[b]	0.39 (1.61)	0.29 (1.53)	0.17	6.00	0.05 [-0.22; 0.32]
	Group[c]	-0.10 (1.63)	-0.45 (1.68)	0.32	3.10	0.17 [-0.13; 0.47]
3.3	Individual	-0.04 (0.47)	0.04 (0.50)	0.31	3.25	-0.16 [-0.43; 0.11]
	Individual[a]	0.01 (0.46)	0.05 (0.50)	0.20	5.14	-0.09 [-0.37; 0.20]
	Group[b]	-0.18 (1.75)	1.71 (1.67)	904715.94	1.11×10^{-6}	-0.94 [-1.28; -0.60]
	Group[c]	-0.55 (1.66)	1.08 (1.46)	1589.85	6.29×10^{-4}	-0.75 [-1.11; -0.40]
3.4	Individual	-0.09 (0.43)	0.09 (0.44)	6.52	0.15	-0.38 [-0.66; -0.11]
	Individual[a]	-0.07 (0.42)	0.10 (0.43)	4.40	0.23	-0.37 [-0.65; -0.08]
	Group[c]	-0.70 (1.76)	2.10 (1.39)	5.43×10^{7}	1.84×10^{-8}	-1.30 [-1.74; -0.87]

[a] *Individual CI trustworthiness ratings excluding participants who misremembered the behaviors and/or had not seen F. Taylor's face.* [b] *Group CI trustworthiness ratings collected after Study 3.3.* [c] *Group CI trustworthiness ratings collected after Study 3.4.*

GENERAL DISCUSSION

We aimed to investigate under which circumstances non-visual behavioral information depicting a person as (un)trustworthy influences one's visual mental representation of that person's facial trustworthiness appearance when one has actually seen the face. Based on Bayesian models of perception, we hypothesized that this would be the case if (1) this type of behavioral information generates a prior about that person's facial trustworthiness appearance (Face Prior Hypothesis) and (2) this face prior is relatively strong compared to sensory input from the actual face (Prior-Likelihood Balance Hypothesis). We also aimed to find out whether behavioral information would have an assimilation or contrast effect on the mentally represented facial appearance. We emphasize that perceived trustworthiness was chosen as example and that we do not claim to have manipulated only perceived trustworthiness and nothing else. The point is whether and under which circumstances non-visual behavioral information biases visual mental representations of seen faces, for which perceived trustworthiness is merely taken as an example in the current studies.

Manipulation checks in all studies showed that the behavioral information successfully manipulated participants' impressions of the target person's trustworthiness. In Study 3.1, we tested and confirmed the Face Prior Hypothesis that behavioral information can generate priors about facial trustworthiness appearance. Behavioral information clearly led to activation of facial features that make the face appear more (un)trustworthy. We subsequently tested the Prior-Likelihood Balance hypothesis by creating conditions of strong sensory input from the face (Study 3.2), weak sensory input (Study 3.3), and strengthened face priors combined with weak sensory input (Study 3.4). The data showed more evidence against an effect on both individual and group CIs in Study 3.2, against an effect on individual CIs but for an assimilation effect on group CIs in Study 3.3, and for an assimilation effect on both individual and group CIs in Study 3.4.

In frequentist terms, group CIs potentially inflate Type I error rates (Jeremy Cone et al., 2020). Therefore, we assign more weight to the findings on the individual CIs in our interpretation. As such, the results suggest that behavioral information has no effect on mental representations when the face is presented

clearly (Study 3.2), still no convincing effect when the face is presented only briefly (Study 3.3), and an assimilation effect when the face prior is mentally visualized before brief presentation of the face (Study 3.4). In summary, the impact of the face prevails, unless face priors are more clearly mentally visualized and less information about the actual face has been encoded.

A Bayesian Inspired View on Social Face Perception

The pattern of results across the four studies seems in line with a standard Bayesian model of perception, which states that sufficiently strong priors yield assimilation effects (Edwards et al., 2012; Kersten & Yuille, 2003). This Bayesian model thus seems useful in informing our hypotheses on *when* and *to what extent* we can expect biases in social face perception. The two predictions we derived from the model specify these circumstances, making them useful informants for hypotheses and experimental designs. The Face Prior Hypothesis implies that if X is to cause a bias in the perception of Y at all, X should generate a prior about Y that can compete with sensory input from Y. The prior and sensory input should thus concern the same object. For the present research, this meant that if non-visual behavioral information were to bias visual mental representations of facial appearance, the behavioral information had to generate a prior about facial appearance. The Prior-Likelihood Balance Hypothesis states that to what extent a bias occurs depends on the balance in strength of the prior and sensory input.

To provide more clarity on the nature of the relative 'strength' of priors and sensory input, we encourage researchers to investigate specific operationalizations for the strength of both. We suggest some possible manipulations here. Regarding the strength of priors, researchers can manipulate the amount and perceived diagnosticity of information underlying the prior (Cone & Ferguson, 2015), reliability of information source (Ernst & Banks, 2002), extent of elaboration, personal relevance of the prior, or a goal to (dis)confirm one's priors (Huang & Bargh, 2014; Kunda, 1990). Regarding the strength of the sensory input, researchers can manipulate completeness and perceived diagnosticity of this information, participants' attentiveness and depth of processing, time passed since information presentation, or a goal to be accurate or not.

Assimilation vs. Contrast

We only found assimilation effects, not contrast effects, in Study 3.3 and 3.4. This is in line with the standard Bayesian model (Edwards et al., 2012; Kersten & Yuille, 2003) and with Mussweiler's (2003) claim that in most comparison situations, people initially focus on similarities, not on differences. Moreover, the results fit with other studies within social face perception that show assimilation effects of earlier information or beliefs on expected or perceived facial appearances (Dotsch et al., 2013; Hassin & Trope, 2000), emotional expression recognition (Bijlstra et al., 2014, 2010; Hugenberg, 2005), and face evaluations (Bliss-Moreau et al., 2008; Falvello et al., 2015). Together, these findings provide accumulating evidence that one's earlier impression of someone can pull one's mental representation of that person's face towards this impression, painting a picture that is more in line with the initial impression. Our current findings do suggest that when it concerns the seen face of a specific individual, such biases are likely to be subtle if they arise at all.

The fact that we found no contrast effects suggests that, in the studies presented here, the actual face was not extremely different from the expected face (Mussweiler, 2003). The actual face was always neutral on trustworthiness, which is perhaps close enough to both an untrustworthy and trustworthy face to allow for assimilation effects. However, contrast effects may still occur under different circumstances. For instance, it remains to be seen whether the current findings generalize to faces that more extremely differ from one's expectations (e.g. a trustworthy person with an untrustworthy looking face).

Implications for Social Face Perception

If opportunity allows, people weigh sensory input from the actual face stronger than their priors of what they expected the face to look like. That is, under circumstances similar to those of our studies. However, when sensory input from the face is so brief that it cannot inform a completely accurate mental representation, chances increase that behavioral information influences the mentally represented face through face priors. This appears to be especially the case when one has a clear prior.

These findings fit with recent studies on the importance of the diagnosticity and reliability of information in informing person evaluations (Brannon &

Gawronski, 2017; Cone & Ferguson, 2015; Cone et al., 2017; Lammers et al., 2017), accounted for by Bayesian models by making probabilities conditional (Zaki, 2013). The findings suggest that, as in everyday cognition (Griffiths & Tenenbaum, 2006) and object perception (Ernst & Banks, 2002; Fetsch et al., 2012), people can weigh social information to inform their impressions in quite an optimal way. Specifically, sensory input from the actual face is generally considered a more reliable cue concerning what the actual face looks like than the face prior based on behavioral information. As the sensory input becomes less reliable due to a brief face presentation time, reliance on the face prior associated with the behavioral information may become somewhat stronger. Given that people often have a clear view of other people's faces, our findings suggest that people's mental representations of others' faces are quite accurate in many situations.

At the same time, the findings of Study 3.4 also suggest that if the prior is sufficiently strong relative to the sensory input, people mentally represent facial features differently from how they actually were. It shows that, although humans can be quite accurate observers, their observation system brings along some biases as well (Edwards et al., 2012; Powers et al., 2017). Although such biases enable humans to observe and act efficiently even in situations with incomplete information, they may lead to significant social consequences (Todorov et al., 2015). Moreover, it may be harder to get rid of inaccurate person evaluations caused by erroneous accusations or gossip, because the evaluation may have leaked through into the mental representation of the face as well. On the other hand, accurate person evaluations based on reliable behavioral information may shield someone with an unfortunate looking face from unfair treatment by biasing the facial appearance to be more in line with the person's behavior. Future studies could investigate this possibility using more extreme looking faces opposed to the neutral face employed in the present studies.

We need to stress though that the effect only emerged convincingly when we combined an explicit instruction to form a face prior with a brief face presentation in Study 3.4. Because the effect has previously been documented when the face was presented to participants for unlimited time (Hassin & Trope, 2000), we had expected the effect to emerge easier. Our current findings suggest, however, that the effect may be less prevalent than extant literature

implies. Moreover, the explicit instruction to form an expectation about the face creates the possibility of a demand effect in Study 3.4. It is conceivable that this instruction gave participants the impression that their expectation of the face was perhaps more important to the study than the briefly presented face, consequently biasing their answers in the RC task in the direction of their face prior. If true, this raises the question to what extent the effect found in Study 3.4 would emerge in real life. Future research may manipulate the visualization of the face prior less obviously. For instance, participants could be asked to mentally visualize a scripted encounter with the person. This increases the chance that they will visualize the facial appearance of the target person without explicitly addressing their attention to their expectation of the face.

The findings from Study 3.4 also question the generalizability of findings from extant literature on social face perception employing the RC task. Study 3.1 showed overwhelming evidence for the generation of a face prior, but the effect of the face prior when the actual face was presented was only found when explicitly instructing participants to mentally visualize the face prior in Study 3.4. This raises the question whether people automatically form an expectation about a person's face or only when prompted to. Although the RC task has proven its ability to visualize mental representations of faces, it is conceivable that effects found with this task only emerge when people are explicitly prompted to think about facial appearance, which they are when doing the RC task. If true, this has important implications for the generalizability of extant and future research employing the RC task.

Last, the relative influence of behavioral information vs. sensory input from the actual face may also depend on the relative decay of memory for behavioral vs. face information. If, for example, the valence of the behaviors sticks more strongly in memory compared to facial details, the effects of behavioral information could be stronger over time than observed in the current studies. This suggestion is open for empirical investigation, where the time between the presentation of the face and the RC task is increased to days or weeks.

Does the Present Research Underestimate the Effect of Behavioral Information on Mentally Represented Faces?

Together, the four studies suggest that people mentally represent others' seen faces quite accurately and that effects of behavioral information tend to be subtle if they arise at all. It is conceivable that influences of behavioral information are even less prevalent in real life than our findings suggest. For instance, behavioral information and the associated face priors may be less extreme, face priors may not even be formed if people are not triggered to think about the face, and people may attend to the face longer than in our studies. However, there are also reasons to suspect that the influence of behavioral information on mentally represented faces may be more prevalent than the current findings suggest.

First, the behavioral information and associated face priors in our studies were perhaps weaker than many in real life. We used only 10 sentences to create an impression about a thus far unknown person. Impressions may often be based on more accumulating information and one may elaborate longer on this information and the impression, making them, and hence their impact on social perception, stronger (Barden & Tormala, 2014). Moreover, people usually tell in detail about a person's behavior in a specific context opposed to reciting short behavioral examples from different scenarios. It is conceivable that one vividly described scenario elicits more emotional and visual processing than 10 short sentences, resulting in a stronger impact as well. Future studies could investigate the effect of established vs. newly formed impressions and of one elaborate scenario vs. multiple one-sentence scenarios.

Second, although perhaps less likely, sensory input from the face may in practice sometimes be even weaker than in our studies. Participants only saw a brief flash of the face in Study 3.3 and 3.4. However, they could still direct full attention to the face, which was presented under otherwise optimal viewing conditions (unobstructed, frontal view under optimal illumination). When meeting others, viewing conditions may be less optimal than in our studies (e.g. shadows on the face, head movements and objects obstructing parts of the face) with distractions in the environment. Future studies could test the same effect under less optimal face viewing conditions.

Third, we measured participants' mental representations immediately after they had seen the face. As mentioned above, one may forget facial details over time (Deffenbacher, Bornstein, McGorty, & Penrod, 2008; Shepherd, Gibling, & Ellis, 1991), while the person impression may remain relatively strong (Bower, 1991; Falvello et al., 2015; Pizarro, Laney, Morris, & Loftus, 2006), allowing it to reshape the faded features. Related to this idea, effects may be stronger when the order of behavioral information and face presentation is reversed. When the face is presented last as in our studies, a clear view of the face allows one to completely overwrite the expected with the real facial appearance. If one sees the face first, one has a clear identity in mind that may then be biased by any new information one learns about the person (Cone, Gunaydin, & DeLong, 2017), especially when details about the actual face begin to fade in memory. Future studies could investigate the effect of time delays and the order of the behavioral and face information.

Fourth, it is conceivable that participants in our studies focused more strongly on remembering the face accurately than people usually do in real life situations. The RC task may feel like a 'test' in which they should try their best to select the faces that look most like the face of the target person. Of course, this is exactly what we ask them to do. Although we tried to subtly decrease this feeling of being tested in Study 3.3 and 3.4, it is likely that participants still interpreted the study in this way. Yet, at present, this 'test' is our best way of measuring and visualizing approximations of their mental representations. As this test does not occur in real life, it is possible that people are slightly less concerned with the accuracy of their mental representations of seen faces in real life, allowing the effect of behavioral information on the mentally represented face to be larger.

A limitation in our studies which may also cause an underestimation of the effect is the possibility of participant demotivation during the long RC task, reducing signal of the mental representation in the CI. Although Study 3.1 provides evidence that participants from the same participation pool can perform a same-length RC task successfully, it is conceivable that it was more tiresome to recreate a (briefly) seen face from memory than an expected face. We did however aim to remove unmotivated participants from the sample through our preregistered exclusion criteria, reducing chances that participant demotivation had a large impact on the current findings.

Rater consistency appeared low, for which several explanations are conceivable. First, repeated ratings (used to calculate consistency but not for main analyses) were given at the end when raters may have become tired and less accurate. Second, the noisiness of CIs may complicate giving consistent ratings. Third, trait ratings vary depending on the distribution of recently encountered faces (Dotsch, Hassin, & Todorov, 2016), which changed with every trial. This may cause low rater consistency but is unlikely to affect our main analyses, because the latter used the average rating across raters, who all rated CIs in random order. Moreover, these average trustworthiness ratings correlated with criterion CI based trustworthiness scores (see Chapter 2), which are not dependent on rater's consistency.

Most issues raised above provide alternative explanations for the strong impact of the actual face on the mental representation of the face. They imply that the impact of the behavioral information on the mentally represented facial appearance after seeing the actual face may be larger in real life than our observed results. In doing so, the raised issues underscore the context dependency of the effect under investigation. Dependent on the relative strength of the prior and sensory input from the actual face, the impact of behavioral information on the mentally represented facial appearance can become larger or smaller.

Conclusion

By combining Bayesian theory with a data-driven methodology, the present research provides novel insight into the circumstances under which mental representations of seen faces are (not) susceptible to biases. The four studies suggest that humans are extraordinarily equipped to perceive faces, mentally representing a person's face quite accurately when they can, only relying on other determinants than the actual face under particular circumstances. Although behavioral information can clearly lead to activation of certain facial appearances, this does not automatically influence mental representations of the seen face of the person who conducted these behaviors. Specifically, the studies suggest that behavioral information only biases mental representations of seen faces in line with the information when it generates strong expectations about the face and sensory input from the actual face is weak. Under these circumstances, the mentally represented facial appearance, known to impact

how people evaluate and interact with the person behind the face, is itself affected by behavioral information about that person. With these results, the present research highlights the context dependency of the effect under investigation. We hope to inspire researchers to further explore circumstances that make the effect (dis)appear opposed to merely showing that the effect exists, thereby increasing our understanding of the conditional probability of such effects in social face perception.

APPENDIX 3A

Behavioral descriptions are taken and translated from Dotsch et al. (2013) and Fuhrman, Bodenhausen, and Lichtenstein (1989). One description was offered by colleagues from our lab group. Descriptions marked with (U) are intended to appear untrustworthy, (N) to appear neutral, and (T) trustworthy.

A pilot study validated that the untrustworthy set (M = -70.91, SD = 39.58) was indeed evaluated as more untrustworthy than the trustworthy set (M = 56.09, SD = 24.97) on a scale ranging from -100 (*untrustworthy*) to 100 (*trustworthy*): δ = -2.54 [-4.03; -1.25], BF_{-0} = 13380.74.

Untrustworthy:

1. Threatened another person with a knife (U)
2. Robbed another person in an alley (U)
3. Provoked a fight while going out (U)
4. Stole another person's jacket at a party (U)
5. Swore at the bus driver (U)
6. Pulled the seat out from underneath somebody (U)
7. Smoked in a non-smoking section even though others complained (U)
8. Tripped someone in the hall (U)
9. Ate a sandwich at the station (N)
10. Crossed the street (N)

Slightly untrustworthy:

1. Shoved a man who was passing out leaflets (U)
2. Pushed another person out of the way to walk through (U)
3. Sneaked in front of people in the queue at the checkout (U)
4. Drank a glass of water (N)
5. Waited at the bus stop (N)
6. Opened the window a little (N)
7. Went grocery shopping at the mall (N)
8. Cleaned the dishes (N)
9. Sent a text message to a friend (N)
10. Had a chat with a neighbor (N)

Slightly trustworthy:

1. Called the neighbor to say that someone tried to break into his car (T)
2. Never smoked in other people's apartments (T)
3. Gave money to charity (T)
4. Took a stroll down the dock (N)
5. Took a can of Coke from the vending machine (N)
6. Locked the door (N)
7. Looked for the bike in the parking (N)
8. Replaced the kitchen light bulb (N)
9. Emptied the trash (N)
10. Turned out the lights before going to bed (N)

Trustworthy:

1. Returned a found wallet (T)
2. Warned a woman that she dropped 50 dollar (T)
3. Volunteered to collect money for poor children (T)
4. Supported a friend in difficult times (T)
5. Lived up to promises made, even if this took a lot of effort (T)
6. Told the cashier that she gave too much change (T)
7. Acted politely and kind to a stranger in the street (T)
8. Helped a man picking up the groceries the man dropped (T)
9. Got an umbrella out as it started to rain (N)
10. Stepped into the elevator (N)

APPENDIX 3B

In the manipulation check, participants were asked to provide their impression of F. Taylor on the following scales:

- Bad – Good
- Incompetent – Competent
- Submissive – Dominant
- Untrustworthy – Trustworthy (scale of interest)
- Unintelligent – Intelligent
- Not criminal – Criminal
- Cold – Warm

The scales ranged from -4 to 4.

APPENDIX 3C

The Face Prior Hypothesis states that verbal information about behavior should generate an expectation about facial appearance (face prior). The hypothesis does not explicitly state whether the influence should run directly from the behavioral information to the face prior (because the behaviors are associated with certain facial features) or (partially) indirectly through a person impression (in this case trustworthiness) that is associated with certain facial features, or yet through a different path. The important point is that, regardless of the exact path, the verbal information about behavior should lead to a visual mental representation of the expected facial appearance.

Attentive readers may have noticed that we implicitly assume the effect to run through the person impression (in this case a trustworthiness evaluation). Indeed, the verbal information about behavior was intended to manipulate participants' beliefs about the trustworthiness of the target person (confirmed in the manipulation check) which we assumed would activate trustworthiness-related facial features. In other words, we assumed the person impression in our research to be both influenced by behavioral information and influencer of the face prior (i.e. a mediator).

Although our goal with the Face Prior Hypothesis is to show that verbal information can generate a visual facial appearance expectation regardless of the exact path through which this happens, readers may wonder whether this assumed mediation effect would hold ground in a statistical mediation analysis. Therefore, although we did not preregister this analysis, we present a mediation analysis testing the mediation effect of verbal information about behavior on the expected facial trustworthiness appearance through the trustworthiness impression of the target person: behavioral information → trustworthiness impression → expected facial trustworthiness appearance.

We conducted a classical mediation analysis in JASP, using the bias-corrected percentile bootstrap method (JASP Team, 2020). We tested whether the effect of behavioral information (0 = untrustworthy; 1 = trustworthy) on the individual CI trustworthiness ratings was mediated by participants' trustworthiness impression of the target person (as measured in the manipulation check). Figure

3C-1 present the results in a path plot. The effect of behavioral information on CI trustworthiness ratings was fully mediated by the trustworthiness impression. The indirect effect was 6.300 * .115 = .725 (p < .001). The bootstrapping procedure computed unstandardized indirect effects for 999 bootstrapped samples, resulting in a 95% confidence interval for the indirect effect of .725 ranging from .358 to 1.184. As the effect of behavioral information on CI trustworthiness ratings was no longer significant when controlling for the trustworthiness impression, the total effect of behavioral information on CI trustworthiness ratings was explained by the trustworthiness impression.

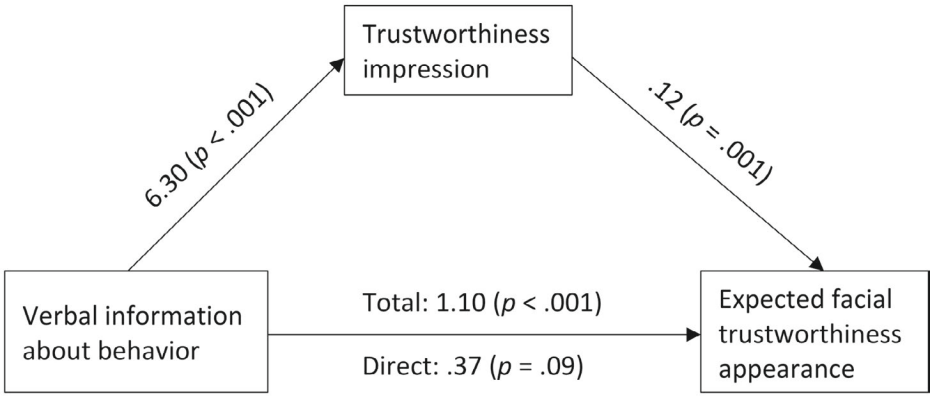


Figure 3C-1. Path plot showing the full mediation of behavioral information on CI trustworthiness ratings by trustworthiness impression.

The results of the mediation analysis show that if the trustworthiness impression indeed is a true mediator of the effect of behavioral information on the expected facial trustworthiness appearance, its mediating impact is significant. That is, our assumption that verbal information about behavior generates a visual expectation about facial appearance because it activates a person impression which is associated with certain facial features is still possibly correct, though certainly not proven. Although the assumption survived the mediation analysis, the mediation analysis in itself cannot guarantee this causal pathway to actually be the right one (Fiedler, Schott, & Meiser, 2011).

While we experimentally manipulated the verbal information, we merely measured both the trustworthiness impression and the mental representation of the expected face, so we cannot make any causal claims regarding the influence

of the trustworthiness impression on the mentally represented expected facial appearance or vice versa. We could in principle swap the mediator and dependent variable in this analysis and state that verbal information about behavior generates a visual expectation about facial appearance (e.g. as scowling/smiling when performing the negative/positive behaviors) which in turn influences the trustworthiness impression of that person. Relatedly, we could argue that both variables reflect the same latent construct (e.g. the target's perceived trustworthiness) measured through different channels (verbal evaluation of the person's trustworthiness and visual facial trustworthiness appearance). Moreover, it is possible that the trustworthiness impression is a spurious mediator if it is correlated with another unmeasured variable which is the true causal mediator. For example, perhaps the behavioral information activates an avoidance/approach tendency or a feeling of rage/delight in the perceiver which is associated with perceiving negative/positive facial appearances. Alternatively, multiple causal pathways may be happening simultaneously (Fiedler et al., 2011).

In conclusion, there are multiple causal pathways conceivable through which the experimentally manipulated behavioral information may influence the visual expectation of facial appearance. The mediation analysis demonstrates that our assumption about the underlying process with person impression as mediator remains a plausible one. As such, a necessary condition for substantiating this assumption is met. Yet, as the analysis in itself does not prove causality, other causal pathways remain possible as well.


4

Temporal Stability of Biases in Mental Representations of Faces

This chapter is based on:
Jansen, L.F., Holland, R.W., Dotsch, R., & Wigboldus, D.H.J. (2021). *Temporal Stability of Biases in Mental Representations of Faces*. Unpublished manuscript. Radboud University, Behavioural Science Institute, Nijmegen, The Netherlands.

The research in this chapter was presented in poster format at the Annual ASPO Conference of the Dutch Association of Social Psychologists on 14 December 2018.

ABSTRACT

Recent research suggests that behavioral information about a person does not bias mental representations of his face after his face is presented for 10 s. However, consequential decisions about people (as in application or eyewitness procedures) often occur hours to months after meeting a person. Are mental representations of faces more likely to be biased the more time has passed since presentation of the face? In Study 4.1 (N=300), we tested whether earlier behavioral information influenced mental representations of a presented face after a time delay of approximately two days. First, participants read behavioral information about an unfamiliar target person depicting him as (un)trustworthy. Participants then saw his face, presented for 10 s. After a delay of approximately two days, participants were asked to remind themselves of their impression of the person. Subsequently, we visualized participants' mental representation of the face using reverse correlation. Behavioral information influenced the mentally represented facial trustworthiness appearance after this time delay. In Study 4.2 (N=300), we repeated the same experiment without time delay. Surprisingly, behavioral information again influenced mental representations. The combined data of both studies provided only partial evidence for a stronger effect of behavioral information after two days compared to immediately after seeing the face. We conclude that, if people remind themselves of their impression of a person, the impression may bias mental representations even of faces recently presented for 10 s. The data hint that this bias could grow stronger over time, but conclusive evidence is lacking. Future research could investigate the effect while increasing the time delay to weeks.

Keywords: social face perception, behavioral information, mental representation, reverse correlation, time delay

Many studies have shown that earlier non-visual information about a person can influence visual representations of that person's face in one's mind (e.g. Brown-Iannuzzi, Dotsch, Cooley, & Payne, 2016; Dotsch, Wigboldus, & Van Knippenberg, 2013; Kunst, Dovidio, & Dotsch, 2017; Ratner, Dotsch, Wigboldus, van Knippenberg, & Amodio, 2014). Participants had not seen the person's face in these studies, so these mental representations concerned participants' expectations of the person's facial appearance. Interestingly, Hassin and Trope (2000) suggested that the influence of non-visual information on visual mental representations of faces occurs even when one has seen the actual face of the target person. In their research, personality information changed the perception of a person's facial features. The way one represents a person's face in one's mind can influence consequential decisions, such as whether to select someone for interrogation, to provide integration support, or to what extent to trust someone (Kunst et al., 2017; Ratner et al., 2014). Therefore, it is important to increase insight into the determinants of such mental representations of faces other than the actual face itself.

In Chapter 3, we extended this literature by investigating the influence of non-visual behavioral information about a target person on visual mental representations of his once seen face. Importantly, instead of merely showing whether the effect exists, we aimed to increase understanding of the circumstances under which the effect (dis)appears. We found convincing evidence for an effect of behavioral information on the mentally represented face only when participants were instructed to mentally visualize their expectation of the face (face prior) and when the actual face was presented for 100 ms (Study 3.4). Evidence was mixed when the actual face was presented for 100 ms and participants were not instructed to visualize their face prior (Study 3.3), speaking for the necessity of visualizing the face prior for the effect to emerge. Moreover, we found evidence *against* the effect when participants were not asked about their face prior, were encouraged to accurately remember the actual face, and when the actual face was presented for 10 s (Study 3.2). This resulted in the suggestion that people will mentally represent a person's face quite accurately when they have had a clear view of the face for 10 s, which is often the case in real life.

However, there are factors conceivable that would decrease the accuracy of the mentally represented face, even if one has seen the actual face for 10 s. One such factor concerns the passing of time after presentation of the face. Although previous research measured mental representations of faces immediately after presentation of the face (Chapter 3; Hassin & Trope, 2000), the moment of seeing the person's face and of making a decision concerning that person often lay apart in time. For instance, application procedures for getting a job, loan, or house usually place the relevant decision hours to weeks after seeing the applicant. Likewise, eyewitness interviews may occur up to weeks or even months after the incident. The moments of encoding and retrieval of the face are thus separated in time. Interestingly, the effect of behavioral information on the mentally represented face may be stronger after such time delays. The reason for this is related to potential differences in memory decay for facial features compared to global person evaluations (i.e. person impressions) based on the behavioral information.

Memory for Perceived Facial Features vs. Global Person Evaluation

In our previous studies, behavioral information triggered a trustworthiness evaluation of the person (as observed in the manipulation check). Such trustworthiness evaluations closely resemble global affective evaluations about a person (Falvello et al., 2015; Oosterhof & Todorov, 2008; Todorov et al., 2008). It is likely that it is easier to remember this trustworthiness or global affective person evaluation over time compared to details about the specific behaviors and facial features (cf. Sanbonmatsu & Fazio, 1990).

Indeed, Förderer and Unkelbach (2013) showed that affective person evaluations remained stable over the course of a week, while memory for the specific sources that triggered such evaluations decreased. They suggest that person evaluations may remain stable over time even without memory for the source. This would imply that even if one does not remember any details about a person, one still remembers one's global evaluation of the person. Similarly, after a delay of 1 week, participants may have biased memory for specific details of a scenario, but still remember whether the person in the scenario acted with good or bad intentions (Pizarro et al., 2006).

In contrast, memory for an unfamiliar face is not so stable. It decreases over time with the steepest decrease shortly after encounter (Deffenbacher et al., 2008). Indeed, facial appearance has already changed in memory after a delay of only 1 day (Ouyang, Hospedales, Song, & Li, 2016). Together, these findings imply that memory for global person evaluations should decay less quickly than for a person's facial features.

Better memory for global person evaluations compared to detailed person information is also in line with theoretical models on evaluations (Bower, 1991; Fazio, 2007). If each piece of detailed information about a person (e.g. behavioral description or facial features) triggers an affective evaluation about the person, this person evaluation becomes stronger and more accessible with each piece of detailed information. After a while, one may have forgotten details about the person, which were presented only once, while still remembering their evaluation about the person, which was rehearsed multiple times. Even if the pieces of detailed information were contradictory in valence, one can integrate such conflicting information into a summary evaluation. From then on, it is likely that one relies on the summary evaluation instead of revisiting the detailed information (Fazio, 2007). One can apply this logic irrespective of the type of theoretical model one chooses to advocate, whether it be associative, connectionist, or propositional (Bower, 1991; De Houwer, 2018; Fazio, 2007).

In conclusion, although people may not remember the exact behavioral and facial details of a person, they may remember their trustworthiness evaluation of the person. Given that trustworthiness is strongly associated with specific facial features (Todorov, 2017; Todorov et al., 2013), the trustworthiness evaluation may influence the facial appearance in the person's mind. As a result, it should become likelier for behavioral information to influence the mental representation of a face presented for 10 s the more time has passed since face presentation.

The Present Research

In the present research, we investigate whether the passing of time after both the behavioral information and the face have been presented strengthens the influence of behavioral information on mentally represented faces. Based on the idea that memory decays more rapidly for facial details compared to

trustworthiness evaluations (informed by the behavioral information), we expect the effect of behavioral information on the mentally represented face to be stronger after a time delay than immediately after seeing the face.

We adopted the same paradigm as in Chapter 3, manipulating trustworthiness evaluations through behavioral descriptions, subsequently presenting the face, and then visualizing mental representations of the face using a reverse correlation (RC) task (Brinkman, Todorov, & Dotsch, 2017; Dotsch & Todorov, 2012; Jack & Schyns, 2017). The resulting classification images (CIs) of the RC task served as approximations of participant's mental representations. We reported ratings of both individual CIs (one CI for each participant) and group CIs (one CI for each experimental group) for each study (see the Methods section for more information). Following the procedure of Study 3.2, in which no effect of behavioral information emerged, we gave participants a clear view of the target person's face for 10 s. Unlike that study, we used three different face identities to prevent that all findings are based on a single face. We also added a novel face selection task for exploratory purposes (see Methods section).

In Study 4.1, we investigated the influence of behavioral information on mental representations of seen faces after a time delay of approximately 2 days. We subsequently conducted Study 4.2, which has no time delay but is in all other ways similar to Study 4.1.[21] We hypothesized that the effect of behavioral information on the mentally represented facial appearance would be notably larger in Study 4.1 than in Study 4.2 (main hypothesis). Moreover, we hypothesized that the effect would be present in Study 4.1 and absent in Study 4.2.

21 The reader may wonder why we chose this 2-Study set up instead of a single study including two time delay conditions (2-day/no time delay). The reason is that we could not reliably predict beforehand whether a time delay of two days would indeed be sufficient for the effect to emerge and that our limited resources excluded the possibility of including a third time delay condition. By starting with the 2-day time delay study, we could have followed with a study including an even larger time delay (instead of no time delay at all) if the effect had not appeared after two days.

STUDY 4.1

We investigated the influence of earlier behavioral information about a person on the mentally represented facial trustworthiness of that person approximately two days after presentation of that person's face. We expected that the mentally represented face would look more trustworthy in the trustworthy than untrustworthy condition.

Method

We preregistered the studies at the Open Science Framework[22] and received ethics approval from the institutional ethics committee. All manipulations, measures, and exclusions are reported. The used paradigm was largely similar to the one used in Chapter 3.

Design and sampling plan

We presented participants with behavioral information about a target person (untrustworthy / trustworthy; between-participants). The degree of facial trustworthiness present in their CIs, as rated by an independent group of raters, formed the dependent variable. In exploratory analyses, we added face identity (Face A / Face B / Face C; between-participants) as independent variable to test whether the effect under investigation generalizes to other face identities than the one used in Chapter 3. Additionally, we ran exploratory analyses with the trustworthiness score of the face selected in a novel face selection task as dependent variable. This short novel task was added towards the end of the study (see below).

As preregistered, we first collected data until 204 participants remained after applying the exclusion criteria. However, some eligible participants did not return on time for the second part of the study after approximately 2 days had passed, leaving us with 199 participants. Figure 4.1 shows the development of the BF for the effect of behavioral information on the individual CI trustworthiness ratings as provided by a first group of raters. As the Bayes factor (BF) for the effect of behavioral information on individual CI trustworthiness was smaller

22 The preregistrations on the Open Science Framework can be found through the following links: Study 4.1 (https://osf.io/f4b5w) and Study 4.2 (https://osf.io/7p486).

than 10 (BF_{-0} < 10 and BF_{0-} < 10),[23] we continued data collection in line with our preregistration until we reached 300 participants after exclusion criteria were applied.[24]

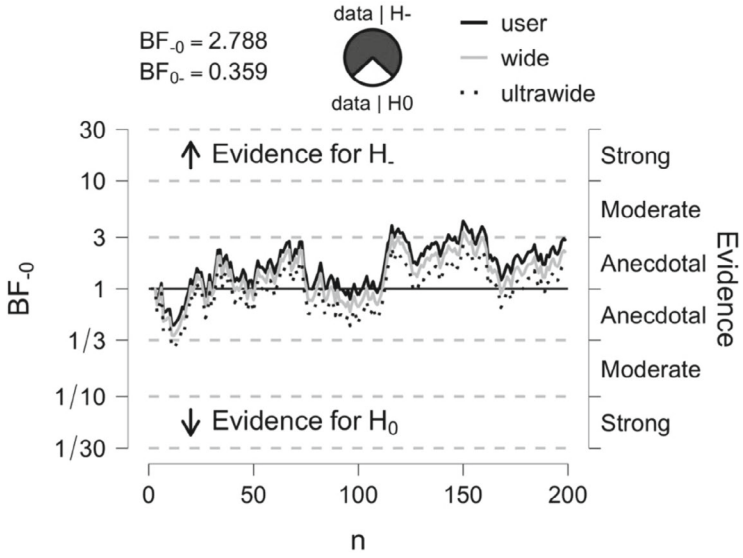


Figure 4.1. Sequential analysis tracking the Bayes factor (BF) for a variety of prior widths (*r*: as used in the analysis / wide / ultrawide) for the effect of behavioral information on individual CI trustworthiness ratings as provided by the first group of raters. The user prior is the JASP default prior. These ratings were used to determine whether to stop data collection (in the case that the BF were at least 10 or 1/10) or to continue until we reached 300 participants.

Participants

We recruited 414 Caucasian adults with normal or corrected-to-normal vision, from a variety of countries and educational backgrounds (see Table S4.1 and S4.2 for participants' countries of residence and levels of education for both studies), on Prolific Academic (https://www.prolific.ac). Following our

23 BFs indicate whether the data are likelier under the null or the alternative model, or whether they do not discriminate the models (Dienes, 2016). A BF of 10 is interpreted as "strong" evidence for one model over the other (Jeffreys, 1961).

24 Because we added two new face identities, we had to create a new base face for the RC task. Consequentially, we could no longer use the criterion CI from Chapter 2 for the sequential analyses in this chapter. Therefore, we returned to the method of having a new group of raters rate the CIs after every so many participants had participated. Although we consider this method less efficient than using the criterion CI, adding the two face identities allowed us to test whether the effect under investigation generalizes to other face identities, which we considered important at this point.

preregistered criteria aimed at excluding unmotivated participants, 99 out of 414 participants were excluded: 37 participants failed the name-behaviors association check (see below), 42 participants failed the attention check (see below), and 20 participants had a median reaction time in the RC task below 500 ms. Moreover, 15 participants never completed the experiment. The final sample included 300 participants (184 women, 114 men, 2 non-binary, M_{age} = 36.31, SD_{age} = 11.76).

Procedure and materials

We designed and hosted the study online using Gorilla (https://gorilla.sc/). We informed participants that the study may end earlier for some participants, who in that case would be paid for their participation up until that point. Moreover, we instructed participants to participate in Part 2 of the study within 44 and 60 hours after completing Part 1. After providing informed consent, participants were instructed to realize a quiet environment without distractions. Subsequently, they performed the following tasks in chronological order.

Behavioral information

Participants were instructed to form impressions about 4 different men (A. Brown, L. Harris, H. Young, and F. Taylor). The men were presented in random order with the exception that F. Taylor, being the target person, was always presented last. Participants read 10 behavioral descriptions about each person (validated in a pilot study; see Appendix 4A) in their own pace with at least 1 s per behavioral description. The descriptions were presented one at a time in random order in the center of the screen underneath the person's name. In the *untrustworthy* condition, F. Taylor was described by the behavioral descriptions from the untrustworthy set. For example, 'F. Taylor threatened another person with a knife'. In the *trustworthy* condition, F. Taylor was described by descriptions from the trustworthy set. For example, 'F. Taylor returned a found wallet'.

Name-behaviors association check

Participants were told that the remainder of the experiment would be about F. Taylor. Participants saw the 4 complete sets of behavioral descriptions presented from left to right in a different order than in the previous task. As memory check,

participants selected the set of behaviors performed by F. Taylor. Participants who selected an incorrect set were excluded from further participation and paid.

F. Taylor's face

As F. Taylor's face, participants saw the grayscale version of either Face A, B or C, which all had a neutral expression (Figure 4.2). Faces A and B were taken from the Chicago Face Database (version 2.0.3; Ma, Correll, & Wittenbrink, 2015). Face A scored higher on trustworthiness than Face B (M = 3.6 vs. M = 3.2), though both can be considered neutral on trustworthiness according to different standards. Face A scored close to the neutral mid-point of the 7-point trustworthiness scale used in the Chicago Face Database. Face B scored equal to the mean and median of the distribution of trustworthiness scores for all White male faces in the Chicago Face Database. Face C was the same neutral face from the Radboud Faces Database (Langner et al., 2010) as used in Chapter 3. F. Taylor's face appeared for 10 s, followed by a mask for 500 ms.

a b c

Figure 4.2. Faces presented as F. Taylor. Faces A and B are taken from the Chicago Face Database version 2.0.3 (Ma et al., 2015). Face C is taken from the Radboud Faces Database (Langner et al., 2010).

Attention check

At the end of Part 1, we included an instructional manipulation check (Oppenheimer et al., 2009). Instead of clicking the continue button to proceed, participants had to click on the title of the instructions page or press the *A* key. Participants who nevertheless clicked on the continue button were excluded from further participation and paid.

Time delay

Participants were told to continue with Part 2 of the study after two days. Specifically, they were told to continue after 44 hours and before 60 hours had passed. The experiment only granted participants access to Part 2 after 44 hours had passed. As access did not automatically close after 60 hours, the time delay was longer than 60 hours for 45 participants.[25] After the delay, participants were again asked to ensure a quiet environment without distractions before starting Part 2.

Refresh memory of target person

To ensure that participants focused on the target person before starting the RC task, they were instructed to think back about who F. Taylor was and what their impression of him was. After 20 s, they could continue with the experiment.

Reverse correlation task

In a two-images forced choice RC task, participants repeatedly selected which of two faces looked most like F. Taylor to them. They did this on 500 trials and could take a break after every 100 trials. Each stimulus consisted of the same base image with random noise superimposed on the image. As base face, we created a grayscale average of the 3 faces that were presented as F. Taylor, smoothened with a Gaussian blur to match the power spectrum of the added noise (Brinkman et al., 2017; Figure 4.3). We generated unique noise patterns for each stimulus, following the procedure described in Dotsch and Todorov (2012). On each trial, the noise patterns for the two stimuli were each other's opposites (Figure 4.3). We generated stimuli in R (R Development Core Team, 2016) using version 0.3.3 of the rcicr package (Dotsch, 2016).

25 Of these 45 participants, 29 participants started Part 2 between approximately 2 and 3 days after Part 1, 6 participants between 3 and 4 days, another 6 participants between 4 and 7 days, and 4 participants between 7 and 11 days. Exploratory Bayesian AN(C)OVA analyses including the interaction effect between behavioral information condition and time delay in hours (or alternatively: time delay longer than 48 or 60 hours (yes/no)) all provided evidence against including the interaction effect, suggesting that the effect of behavioral information did not change remarkably over time within Study 4.1.

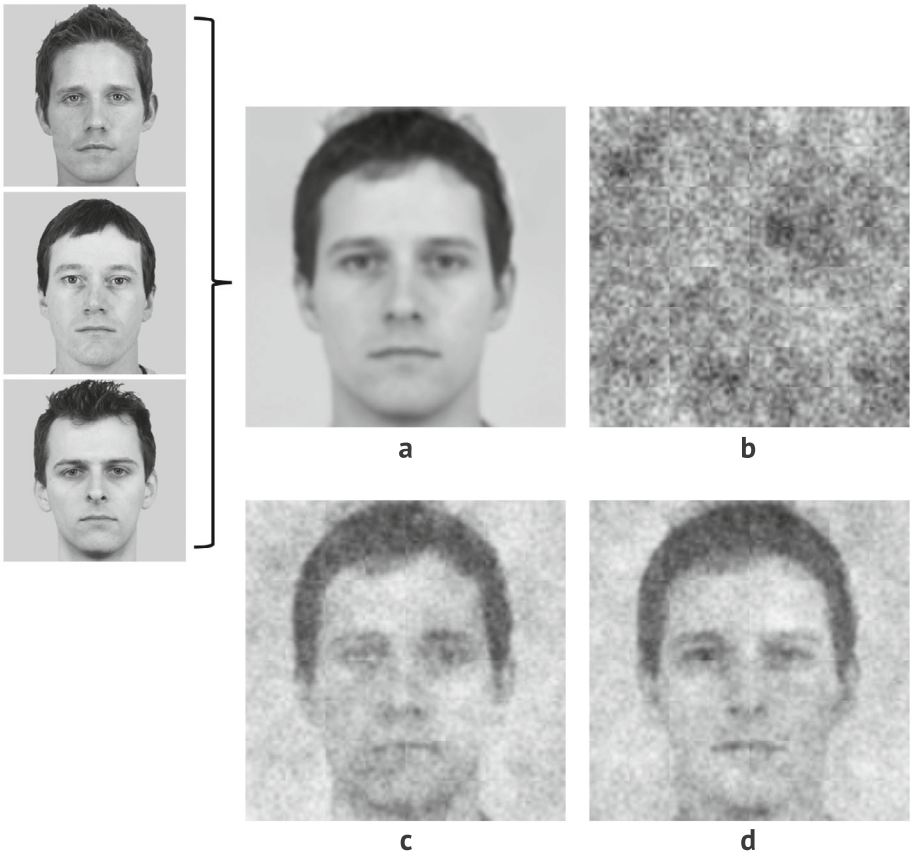


Figure 4.3. (a) Base face used in the reverse correlation task. The face is the average of the three faces presented as F. Taylor (shown left), smoothened with a Gaussian blur to match the power spectrum of the added noise in the reverse correlation task (Brinkman et al., 2017). (b) Example of a random noise pattern. (c) Example stimulus of the base face with a random noise pattern added and (d) with its inverse pattern added.

Manipulation check

Participants evaluated F. Taylor on a trustworthiness scale, ranging from -4 (*untrustworthy*) to 4 (*trustworthy*). To disguise our interest in trustworthiness, we embedded the scale in other scales (Appendix 4B).

Face selection task

As an explorative measure of participants' mental representations, we included a novel face selection task much shorter than the RC task. In this task, participants selected the face they believed was F. Taylor from a set of 5 faces (Figure 4.4). The set contained the face originally presented as F. Taylor to the

participant as well as transformed versions of the face appearing increasingly untrustworthy and trustworthy. The original face was presented left in the bottom row. We transformed the faces in WebMorph (DeBruine, 2017), with continuum settings set to 40% for shape and texture and 0% for color in 3 steps, resulting in 7 images. As untrustworthy and trustworthy faces for the transform dimension, we used the -3 SD and +3 SD versions of face identity 24 from the face database '25 White Faces Manipulated on Trustworthiness' generated using FaceGen Modeller 3.2 (http://facegen.com) as described by Todorov and colleagues (Todorov et al., 2013; Todorov & Oosterhof, 2011). We excluded the 2 extreme faces from the set of 7 images, resulting in the final set of 5 images, which ranged from -2 (*untrustworthy*) to 2 (*trustworthy*), with the original face scoring 0. The 5 images were presented on screen until participants made a selection.

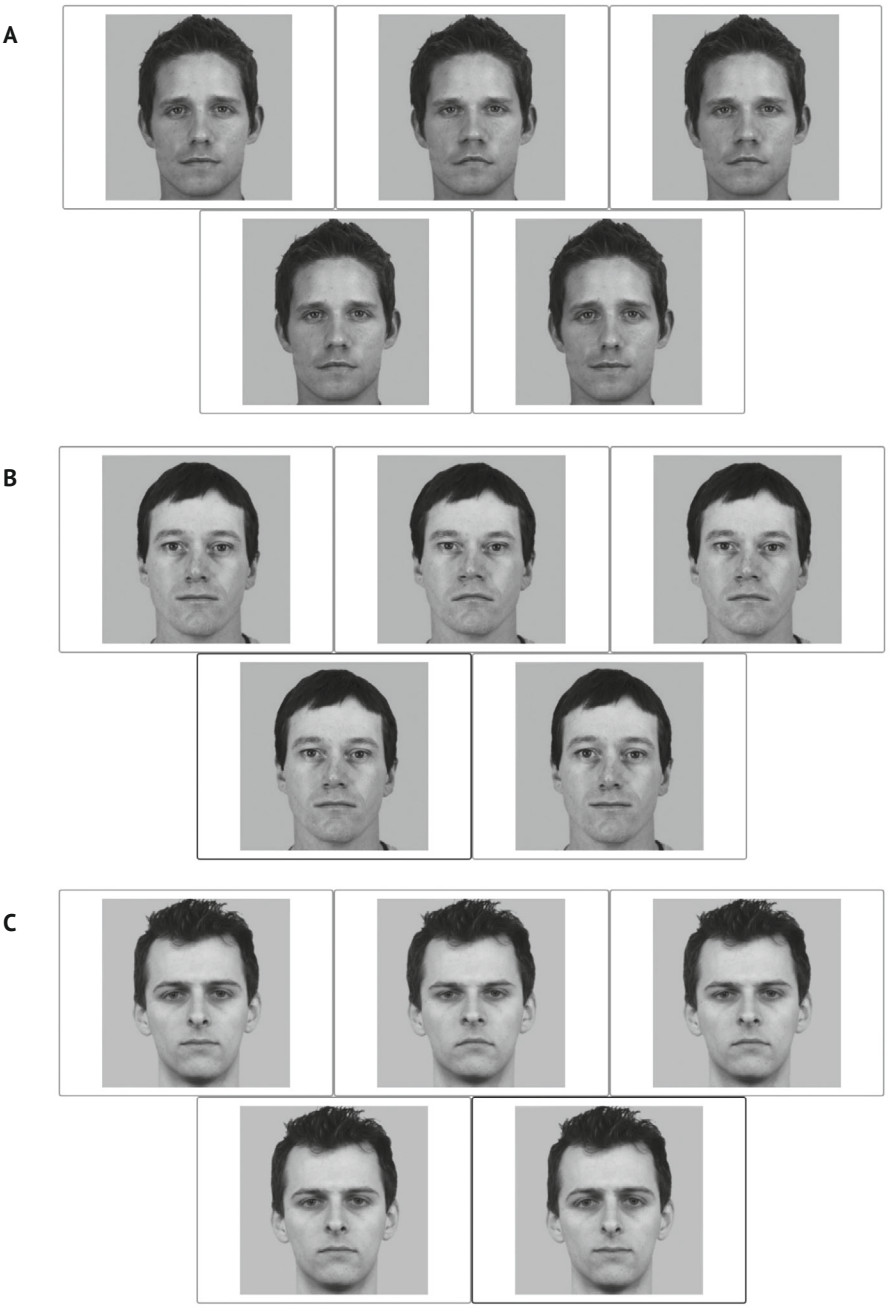


Figure 4.4. Sets of 5 faces varying on trustworthiness for the face selection task. Participants only saw the set corresponding to the face identity they had seen as F. Taylor.

Questionnaire

Participants provided answers about their experience during the experiment (see Appendix S-4A for the questions), demographics (gender, age, native language, country of residence, and highest completed education), and proficiency in English. They could also leave any remarks. Last, participants were sent back to Prolific and paid.

CI generation and trustworthiness evaluation

CI generation

We used version 0.4.0 of the rcicr package (Dotsch, 2017) to generate CIs. For each participant, we averaged all selected noise patterns to generate the individual classification patterns (CPs) for that participant. By scaling and superimposing CPs on the base image, we generated the individual classification images (CIs), serving as approximate visualizations of participants' mental representations of F. Taylor. We computed group CPs by averaging the raw individual CPs, which were then also visualized as group CIs, per behavioral information condition (untrustworthy / trustworthy) collapsed over face identities (Figure 4.5) and for each face identity separately (Figure 4.6).

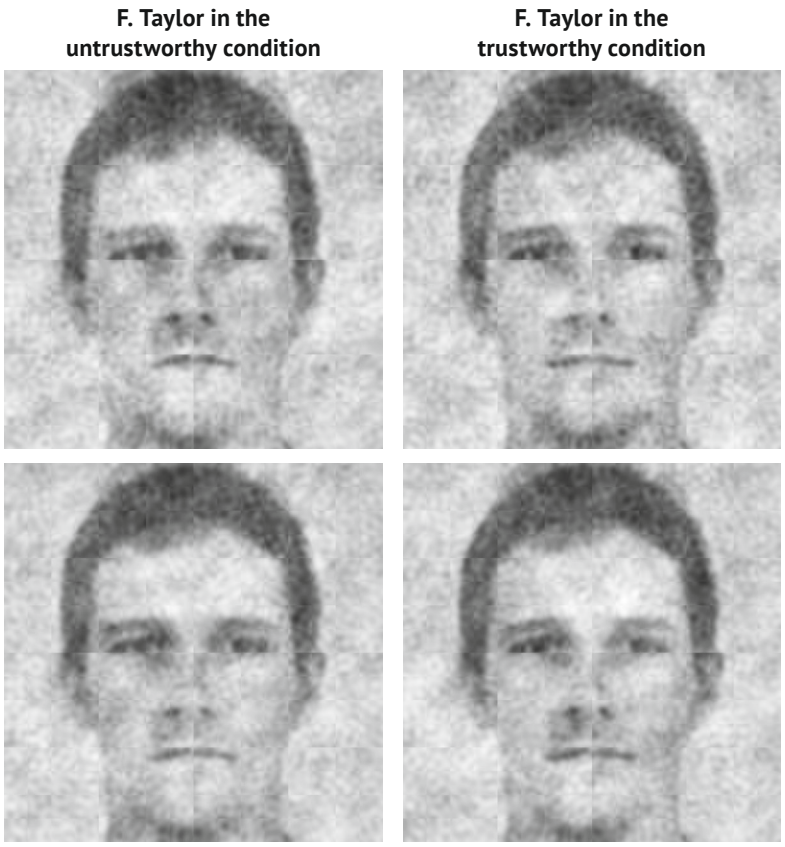


Figure 4.5. Visualization of the group mental representations of F. Taylor's facial appearance when F. Taylor was described as an untrustworthy (left column) or trustworthy (right column) person and mental representations were assessed at least 2 days (Study 4.1; top row) or immediately (Study 4.2, bottom row) after presenting F. Taylor's face.

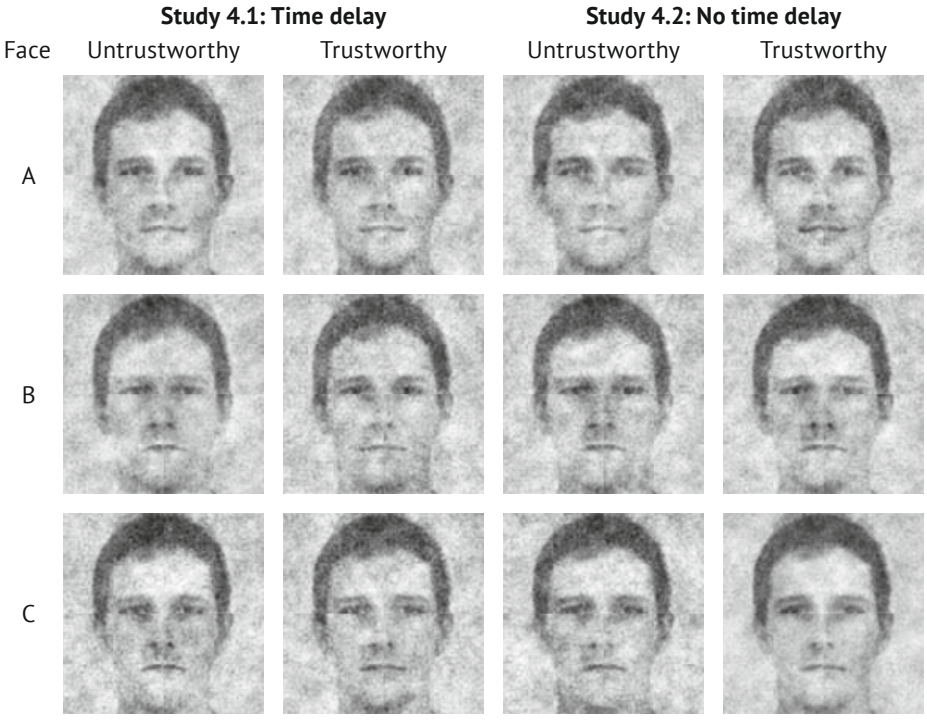


Figure 4.6. Visualization of the group mental representations of F. Taylor's facial appearance in Study 4.1 (time delay, first 2 columns) and Study 4.2 (no time delay, last 2 columns) when he was described as an untrustworthy (Columns 1 and 3) or trustworthy (Columns 2 and 4) person and looked like face A (top row), B (middle row), or C (bottom row).

Individual CIs rating task

An independent group of 22 Caucasian raters evaluated the 300 individual CIs on a 9-point scale ranging from -4 (*very untrustworthy*) to 4 (*very trustworthy*). The CIs were presented in random order. Following our preregistered exclusion criteria, 2 raters were excluded because their median reaction time was below 500 ms. The final sample included 20 raters (13 women, 7 men, M_{age} = 32.40, SD_{age} = 14.58, $M_{consistency}$ = .34, $SD_{consistency}$ = .20).[26] Raters evaluated an a priori selected subset of 20 CIs again, without knowing that these were repeated trials. We computed the intraclass correlation ($ICC_{3,1}$; Shrout & Fleiss, 1979)

26 The first 199 individual CIs were previously evaluated by another independent group of 20 raters (16 women, 4 men, M_{age} = 33.60, SD_{age} = 12.16, $M_{consistency}$ = .27, $SD_{consistency}$ = .20). At that time, the data provided weak evidence for an effect of behavioral information on CI trustworthiness ($M_{untrustworthy}$ = -0.07, $SD_{untrustworthy}$ = 0.42 vs. $M_{trustworthy}$ = 0.08, $SD_{trustworthy}$ = 0.56, δ = -.29 [-.57; -.05], BF_{-0} = 2.79). As the BF was below 10, we continued data collection as preregistered until we reached 300 participants.

between a rater's initial and repeated judgment of these CIs as indication of that rater's consistency.[27] The initial ratings were used to score the individual CIs on perceived trustworthiness. After standardizing initial ratings per rater,[28] the average trustworthiness rating for each individual CI was computed.

Group CIs forced choice and rating task

Another independent group of 40 Caucasian raters (26 women, 14 men, M_{age} = 30.83, SD_{age} = 11.66) evaluated the group CIs of the untrustworthy and trustworthy conditions collapsed over face identities, as well as per face identity. As forced choice task, participants selected for each face identity (collapsed / A / B / C) which of the two group CIs (untrustworthy / trustworthy) looked most trustworthy to them. As rating task, participants rated each of the 8 group CIs on a 9-point scale ranging from -4 (*very untrustworthy*) to 4 (*very trustworthy*).

Results[29]

Confirmatory analyses

As preregistered, we conducted Bayesian analyses using the default prior settings in JASP (JASP Team, 2020) in all studies. Each resulting BF represents a continuous measure of evidence for one model over another model. The 95% credible intervals of estimated effect sizes are reported between brackets. Bayesian one-sided *t*-tests looked into the effect of behavioral information (untrustworthy / trustworthy) collapsed over face identities. The one-sided hypothesis that the untrustworthy condition scored lower on trustworthiness than the trustworthy condition was represented by a Cauchy prior distribution with $r = 1/\sqrt{2}$, truncated to allow negative effect size values only.

Manipulation check

As predicted, participants evaluated F. Taylor as less trustworthy in the untrustworthy (M = -2.55, SD = 1.79) than trustworthy condition (M = 2.20,

27 We did not specify the type of correlation in our preregistration. We chose to compute intraclass correlations Model 3, Form 1, because it is regarded to be a good indicator of intrarater consistency (Shrout & Fleiss, 1979; Trevethan, 2017).

28 Using standardized or unstandardized ratings did not change the conclusions for any of the studies.

29 In both studies, the assumption of normality appeared to be violated in some analyses. Whenever we could, we corrected for this (through data transformation or conduction of a non-parametric test) and reported results of these corrected analyses in the main text. Such corrections never changed the conclusions of the analyses.

SD = 1.83, posterior median effect size δ = -2.60, 95% credible interval [-2.93; -2.29], BF_{-0} = 3.37 x 10^{63}). Due to violation of the assumption of normality, a Bayesian Mann-Whitney U test was conducted to test the result's robustness. The conclusion remained the same (BF_{-0} = 1.06 x 10^{11}).

Individual CI trustworthiness

Because the assumption of normality was violated, we transformed the dependent variable by adding 2 (to get all positive values) and subsequently taking the square root. For the sake of interpretation, we report the means and standard deviations of the untransformed ratings. As predicted, individual CIs of F. Taylor were rated lower on trustworthiness in the untrustworthy (*M* = -.10, *SD* = .43) than trustworthy condition (*M* = .10, *SD* = .47, δ = -.42 [-.65; -.20], BF_{-0} = 231.40).[30] Figure 4.7 shows the development of the BF as sample size increases.

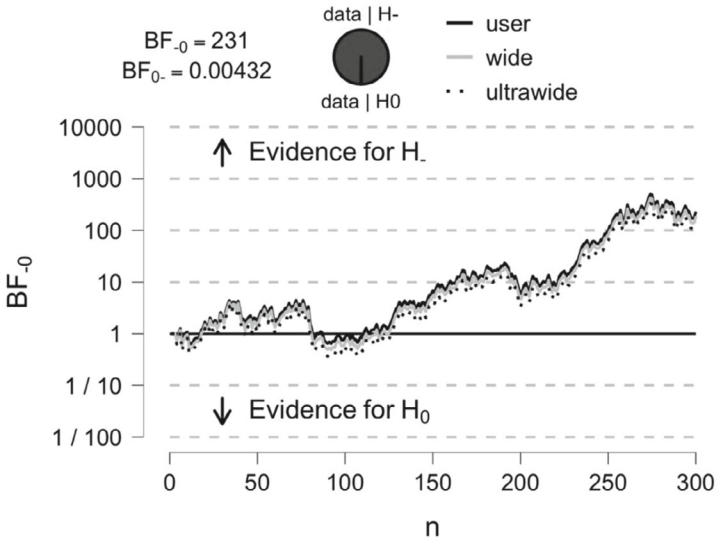


Figure 4.7. Sequential analysis tracking the Bayes factor (BF) for a variety of prior widths (*r*: as used in the analysis / wide / ultrawide) for the effect of behavioral information on individual CI trustworthiness ratings as sample size increases in Study 4.1. The user prior is the JASP default prior.

30 It is possible that 27 participants may have misremembered which behaviors were performed by F. Taylor by the time of the manipulation check. Their ratings on this task were opposite in valence to the behavioral information they read earlier in the study. Excluding these participants did not change the conclusions (BF_{-0} = 282.98).

Group CI trustworthiness

As predicted, the group CI from the trustworthy condition was selected as appearing most trustworthy by 36 out of 40 raters, BF_{10} Independent multinomial = 355004.53. Also, as predicted, group CIs of F. Taylor were rated as less trustworthy in the untrustworthy (*M* = -1.18, *SD* = 1.38) than trustworthy condition (*M* = 1.65, *SD* = 1.21, δ = -1.71 [-2.22; -1.22], $BF_{-0} = 1.48 \times 10^{11}$). Due to a violation of normality the result's robustness was tested with a Bayesian Wilcoxon signed-rank test, which resulted in the same conclusion ($BF_{-0} = 1.39 \times 10^{6}$).

Exploratory analyses

Face selection task

The data showed moderate evidence for an effect of behavioral information on the perceived trustworthiness of the face selected in the face selection task. They suggest that participants selected a less trustworthy looking face in the untrustworthy (*M* = -0.24, *SD* = 1.13) than trustworthy condition (*M* = 0.07, *SD* = 1.10, δ = -0.27 [-0.49; -.06], BF_{-0} = 3.92).

Face identity effects

We added face identity as independent variable to the confirmatory analyses described above. Reported Inclusion BFs ($BF_{Inclusion}$) compare the models including the interaction effect with the equivalent models excluding the interaction effect. For each analysis, we used the default prior settings in JASP (JASP Team, 2020). For the group CI forced choice data, we created the Bayesian contingency tables for each face identity (A / B / C) separately.

Manipulation check

No effects of or with face identity were supported (all $BF_{Inclusion}$s < 0.10).

Individual CI trustworthiness

Both main effects of behavioral information (BF_{10} = 115.70) and face identity (BF_{10} = 15926.79) were convincing in a 2 (behavioral information: untrustworthy / trustworthy) x 3 (face identity: Face A / Face B / Face C) Bayesian ANOVA. Post hoc tests revealed that individual CIs of Face A (*M* = 0.19, *SD* = 0.48) looked more trustworthy than those of Face B (*M* = -0.09, *SD* = 0.42, $BF_{10, U}$ = 1357.00) and C

(M = -0.11, SD = 0.42, $BF_{10, U}$ = 3935.81). The data did not support a Behavioral Information x Face Identity interaction effect ($BF_{Inclusion}$ = 0.08).[31]

Group CI trustworthiness

On the forced choice trials, the group CI from the trustworthy condition was selected as appearing most trustworthy by 24 out of 40 participants for Face A (BF_{01} Independent multinomial = 1.24), 38 out of 40 participants for Face B (BF_{10} Independent multinomial = 3.20 x 10^{7}), and 36 out of 40 participants for Face C (BF_{10} Independent multinomial = 484097.09).

The data from the rating task supported inclusion of both main effects of behavioral information (BF_{10} = 2.35 x 10^{14}) and face identity (BF_{10} = 7.95 x 10^{14}), as well as the Behavioral Information x Face Identity interaction effect ($BF_{Inclusion}$ = 1.21 x 10^{7}) in a Bayesian repeated measures ANOVA. Post hoc tests showed that group CIs of Face A looked more trustworthy than those of Face B ($BF_{10, U}$ = 7.16 x 10^{13}) and C ($BF_{10, U}$ = 6.85 x 10^{11}). Follow-up paired samples t-tests showed that the effect of behavioral information was present for each face identity, but that it was much smaller for Face A ($M_{untrustworthy}$ = 1.70, SD = 1.20 vs. $M_{trustworthy}$ = 2.25, SD = 1.32, δ = -.40 [-.72; -.10], BF_{-0} = 8.28), than for Face B ($M_{untrustworthy}$ = -1.88, SD = 1.22 vs. $M_{trustworthy}$ = 1.23, SD = 1.25, δ = -1.76 [-2.28; -1.26], BF_{-0} = 3.34 x 10^{11}) and C ($M_{untrustworthy}$ = -1.78, SD = 1.78 vs. $M_{trustworthy}$ = 0.95, SD = 1.36, δ = -1.01 [-1.41; -.63], BF_{-0} = 512730.52). Wilcoxon signed rank tests lead to similar conclusions (Face A: BF_{-0} = 21.47; Face B: BF_{-0} = 129835.10; Face C: BF_{-0} = 36919.39).

Discussion

Study 4.1 investigated the effect of behavioral information on the mentally represented facial trustworthiness appearance after a time delay. Presentation of F. Taylor's face lasted for 10 s, allowing participants to take a good look at the face. This set-up largely resembles the procedure of Study 3.2, which provided evidence against an effect of behavioral information. With the addition of a time delay of approximately two days in the current study, we hypothesized that behavioral information would bias the mentally represented facial trustworthiness appearance in line with the information. Indeed, this is what we found.

31 Which is equal to 1 / 0.06 = 16.67 against including the interaction effect.

It is important to note, however, that Study 4.1 differed from Study 3.2 in subtle other ways than time delay only. First, Study 4.1 used two other face identities next to the face identity used in Chapter 3. Second, the specific instructions used for face presentation differ between the two studies (memory focused vs. no specific focus). Third, participants were instructed to remind themselves of their impression of the person before visualizing their mental representation of his face in Study 4.1. Fourth, the 10 s face presentation in Study 3.2 was not followed by a mask. These differences could serve as alternative explanations for why the effect emerged in Study 4.1, but not in Study 3.2. To eliminate these subtle differences, we conducted Study 4.2.

STUDY 4.2

Study 4.2 is an exact replication of Study 4.1, except that it contains no time delay. Based on Study 3.2, we hypothesized that the effect of behavioral information on mentally represented facial trustworthiness appearance would be absent in Study 4.2. Moreover, we hypothesized that the effect would be larger in Study 4.1 (with a time delay) than Study 4.2 (without a time delay; main hypothesis).

Method

Except where indicated below, the method was identical to Study 4.1.

Sampling plan

To match the cell sizes of Study 4.1, we collected data of 300 participants after exclusion criteria were applied.

Participants

Following the preregistered criteria, 109 out of 409 participants were excluded: 40 participants failed the name-behaviors association check, 32 failed the attention check, and 37 had a median reaction time in the RC task below 500 ms. The final sample included 300 participants (183 women, 117 men, M_{age} = 37.88, SD_{age} = 12.26).

Materials and procedure

No time delay

After Part 1, participants were told that they would now continue with Part 2 of the study and were immediately granted access to it.

CI generation and trustworthiness evaluation

Individual CI rating task

One rater was excluded because his median reaction time was below 500 ms. The final sample included 20 raters (14 women, 6 men, M_{age} = 30.80, SD_{age} = 8.20, $M_{consistency}$ = .44, $SD_{consistency}$ = .21).

Group CIs forced choice and rating task

Another independent group of 40 Caucasian raters (19 women, 21 men, M_{age} = 34.53, SD_{age} = 13.80) evaluated the group CIs of Study 4.2.

Results

We first present the results for the analyses of Study 4.2, testing our prediction of more evidence against an effect of behavioral information on the mental representation. Hereafter, we present the results for the analyses on the combined data of Study 4.1 and 4.2, testing our prediction that the effect of behavioral information should be larger with a time delay (Study 4.1) than without a time delay (Study 4.2; main hypothesis).

Confirmatory analyses Study 4.2

The conducted analyses were identical to Study 4.1, except that the *t*-tests on the CI trustworthiness ratings were two-sided as we expected more evidence for the null model in those analyses. In those cases, the Cauchy prior distribution with $r = 1/\sqrt{2}$ allowed for both negative and positive effect size values.

Manipulation check

As predicted, participants evaluated F. Taylor as less trustworthy in the untrustworthy (M = -3.06, SD = 1.55) than trustworthy condition (M = 3.18, SD = 1.48, δ = -4.10 [-4.44; -3.76], BF_{-0} = 5.58 x 10^{105}), which was also found in a Bayesian Mann-Whitney U test (BF_{-0} = 1.18 x 10^{12}).

Individual CI trustworthiness

Because the assumption of normality was violated, we transformed the ratings in the same way as in Study 4.1. We again report means and standard deviations of the untransformed ratings for ease of interpretation. Contrary to our expectation, individual CIs of F. Taylor were rated lower in the untrustworthy (M = -.11, SD = .52) than trustworthy condition (M = .11, SD = .50, δ = -.41 [-.64; -.19], BF_{10} = 85.32).[32] Figure 4.8 shows the development of the BF as sample size increases.

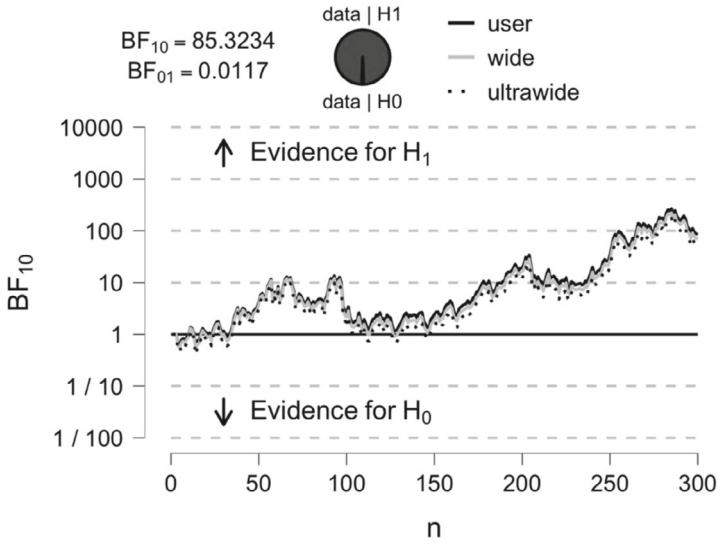


Figure 4.8. Sequential analysis tracking the Bayes factor (BF) for a variety of prior widths (*r*: as used in the analysis / wide / ultrawide) for the effect of behavioral information on individual CI trustworthiness ratings as sample size increases in Study 4.2. The user prior is the JASP default prior.

Group CI trustworthiness

On the forced choice task, the group CI of the trustworthy condition was selected as most trustworthy by 26 out of 40 raters, BF_{10} independent multinomial = 2.24, which is slightly more than the approximate 20 raters we expected. Contrary to our expectations, the group CI of F. Taylor was rated as less trustworthy in

32 Thirteen participants may have misremembered which behaviors were performed by F. Taylor by the time of the manipulation check. Their ratings on this task were opposite in valence to the behavioral information presented to them. Excluding these participants did not change the conclusions (BF_{10} = 53.83).

the untrustworthy (*M* = -0.53, *SD* = 1.34) than trustworthy condition (*M* = 1.50, *SD* = 1.18, δ = -1.26 [-1.70; -.84], BF_{10} = 2.82 x 10^7).

Exploratory analyses Study 4.2

Face selection task

Data from the face selection task provided weak support for the null hypothesis, suggesting that the faces selected in the untrustworthy condition (*M* = -0.18, *SD* = 1.08) did not differ much on trustworthiness from those selected in the trustworthy condition (*M* = 0.01, *SD* = 1.07, δ = -.17 [-.39; .06], BF_{01} = 2.70).

Face identity effects

Manipulation check

Again, no effects of or with face identity were supported (all $BF_{Inclusion}$s < 0.09).

Individual CI trustworthiness

As in Study 4.1, both main effects of behavioral information (BF_{10} = 85.32) and face identity (BF_{10} = 149.44) were convincing in the 2 (behavioral information: untrustworthy / trustworthy) x 3 (face identity: Face A / Face B / Face C) Bayesian ANOVA. Here too, post hoc tests revealed that individual CIs of Face A (*M* = 0.18, *SD* = 0.54) looked more trustworthy than those of Face B (*M* = -0.12, *SD* = 0.46, $BF_{10, U}$ = 394.16) and C (*M* = -0.05, *SD* = 0.52, $BF_{10, U}$ = 12.40). Again, the data did not support a Behavioral Information x Face Identity interaction effect ($BF_{Inclusion}$ = 0.15).[33]

Group CI trustworthiness

On the forced choice trials, the group CI from the trustworthy condition was selected as appearing most trustworthy by 31 out of 40 participants for Face A (BF_{10} Independent multinomial = 162.30), 23 out of 40 participants for Face B (BF_{01} Independent multinomial = 1.77), and 34 out of 40 participants for Face C (BF_{10} Independent multinomial = 8465.04).

The data from the group CIs rating task of Study 4.2 supported inclusion of both main effects of behavioral information (BF_{10} = 4765.32) and face identity

33 Which is equal to 1 / 0.15 = 6.67 against including the interaction effect.

(BF_{10} = 6.66 x 10^{34}), but not of the Behavioral Information x Face Identity interaction effect ($BF_{Inclusion}$ = 0.13).[34] Post hoc tests showed that group CIs of all three faces differed from each other on trustworthiness with Face A looking more trustworthy than Face B ($BF_{10, U}$ = 2.39 x 10^{25}) and C ($BF_{10, U}$ = 2.38 x 10^{13}), and Face B looking less trustworthy than Face C ($BF_{10, U}$ = 4.52 x 10^{12}). As on the individual CI ratings of Study 4.2, the effect of behavioral information was approximately equally strong for each face identity.

Interim conclusion Study 4.2

Without the time delay in Study 4.2, we expected to find more evidence for the null hypothesis, as in Study 3.2. However, the data provided more evidence for than against an effect of behavioral information on the mentally represented facial appearance. It appears that, even without a time delay, behavioral information can bias the mentally represented facial trustworthiness appearance of a face presented for 10 s in line with the behavioral information. Yet, it is still possible that the effect of behavioral information is larger after a time delay (Study 4.1) than immediately after seeing the face (Study 4.2), as outlined in our main hypothesis. As preregistered, we combined the data of both studies into one dataset to statistically test for this possibility.

Confirmatory analyses combined dataset

By combining the data from Study 4.1 and 4.2 into a single dataset, we could add time delay (yes / no; between-participants) as independent variable. We conducted 2 (behavioral information: untrustworthy / trustworthy) x 2 (time delay: yes / no) Bayesian (repeated measures) ANOVAs with default prior settings in JASP (JASP Team, 2020).

Manipulation check

Besides an overall main effect of behavioral information (BF_{10} = 5.37 x 10^{163}), we found that the effect of our manipulation was bigger at the time of the manipulation check in Study 4.2 (without time delay) than in Study 4.1 (with time delay; $BF_{Inclusion}$ = 195459.49). This suggests that the manipulation of the trustworthiness impression became somewhat weaker over time.

34 Which is equal to 1 / 0.13 = 7.69 against including the interaction effect.

Individual CI trustworthiness

We again transformed the ratings the same way as in Study 4.1. The combined data supported a main effect of behavioral information (BF_{10} = 68089.64). Contrary to our main hypothesis, we found evidence against a Behavioral Information x Time Delay interaction effect ($BF_{Inclusion}$ = 0.14).[35] The effect of behavioral information on individual CI trustworthiness ratings was thus approximately equally strong with and without time delay.[36]

Group CI trustworthiness

The combined data provided evidence for a main effect of behavioral information (BF_{10} = 4.28 x 10^{23}) and weak evidence for an interaction effect ($BF_{Inclusion}$ = 2.01). The weak interaction effect suggests that the effect of behavioral information was larger after a time delay than after no time delay (Figure 4.9), which is in line with the main hypothesis.

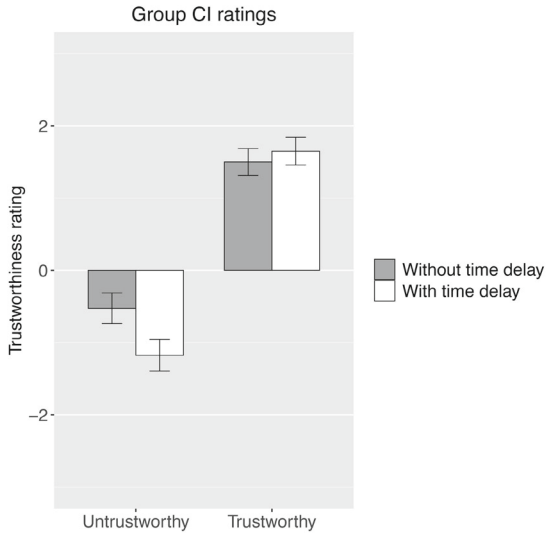


Figure 4.9. Average trustworthiness ratings of group CIs of F. Taylor when participants created the CI immediately (Study 4.2: without time delay) or at least 2 days (Study 4.1: with time delay) after seeing F. Taylor's face.

35 Which is equal to 1 / 0.14 = 7.14 against including the interaction effect. An exploratory Bayesian ANCOVA with time delay indicated in hours (opposed to yes/no) yielded similar results ($BF_{Inclusion}$ = 0.18).

36 Excluding the participants from Study 4.1 and 4.2 that later may have misremembered the behaviors linked with F. Taylor did not change any of the conclusions (BF_{10} = 45492.98 for the effect of behavioral information and $BF_{Inclusion}$ = 0.13 for the interaction effect with time delay).

Exploratory analyses combined dataset

Face selection task

Although the separate data of Study 4.1 and 4.2 provided more evidence against and for the null hypothesis respectively, the combined data of Study 4.1 and 4.2 provided no support for an interaction effect ($BF_{Inclusion} = 0.16$).[37] They provided moderate support for a main effect of behavioral information ($BF_{10} = 3.84$) only.

Face identity effects

We again added face identity as independent variable to the confirmatory analyses described above.

Manipulation check

Again, no effects of or with face identity were supported (all $BF_{Inclusion}$s < 0.08).

Individual CI trustworthiness

We again used the transformed ratings and report means and standard deviations from the untransformed ratings for clarity. A 2 (behavioral information: untrustworthy / trustworthy) x 2 (time delay: yes / no) x 3 (face identity: Face A / Face B / Face C) Bayesian ANOVA with default prior settings provided evidence only for main effects of behavioral information ($BF_{10} = 68089.64$) and face identity ($BF_{10} = 2.58 \times 10^7$). Though the pattern of results hinted at a three-way interaction (Figure 4.10), we found no evidence for any of the interaction effects (all $BF_{Inclusion}$s < 0.26). With and without time delay, individual CIs of F. Taylor looked more trustworthy in the trustworthy ($M = 0.10$, $SD = 0.49$) than untrustworthy condition ($M = -0.10$, $SD = 0.47$), and more trustworthy for Face A ($M = 0.18$, $SD = 0.51$) than for Face B ($M = -0.11$, $SD = 0.44$, $BF_{10, U} = 3.54 \times 10^6$) and C ($M = -0.08$, $SD = 0.47$, $BF_{10, U} = 111325.95$).

37 Which is equal to 1 / 0.16 = 6.25 against including the interaction effect.

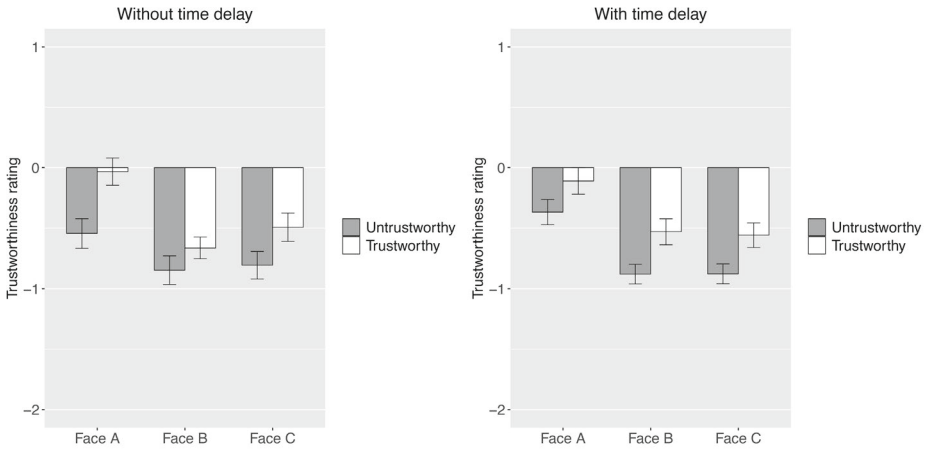


Figure 4.10. Average trustworthiness ratings of **individual CIs** of F. Taylor per behavioral information (untrustworthy or trustworthy), face identity (Face A, B, or C), and time delay (without or with time delay) condition. For ease of interpretation, untransformed unstandardized ratings are shown.

Group CI trustworthiness

A 2 (behavioral information: untrustworthy / trustworthy) x 2 (time delay: yes / no) x 3 (face identity: Face A / Face B / Face C) Bayesian repeated measures ANOVA with default prior settings provided evidence for main effects of behavioral information (BF_{10} = 9.37 x 10^{16}) and face identity (BF_{10} = 1.05 x 10^{43}), as well as for all interaction effects (all $BF_{Inclusion}$s > 375.93). We performed follow-up 2 (behavioral information: untrustworthy / trustworthy) x 2 (time delay: yes / no) Bayesian repeated measures ANOVAs for each face identity separately. In line with our main hypothesis, the effect of behavioral information on the group CI trustworthiness ratings was indeed bigger with time delay than without time delay for Faces B ($BF_{Inclusion}$ = 71130.63) and C ($BF_{Inclusion}$ = 14.97). However, contrary to our main hypothesis, data for Face A showed very weak evidence for the effect in the reversed direction ($BF_{Inclusion}$ = 1.43). Figure 4.11 displays this three-way interaction.

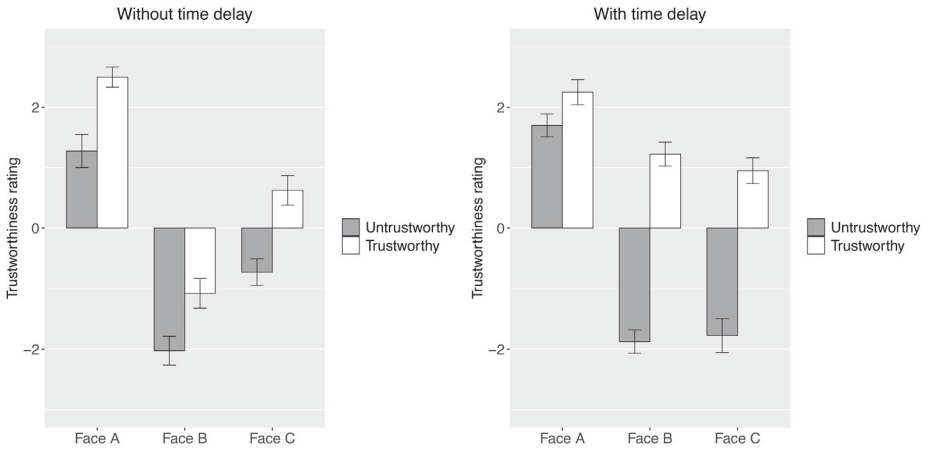


Figure 4.11. Average trustworthiness ratings of **group CIs** of F. Taylor per behavioral information (untrustworthy or trustworthy), face identity (Face A, B, or C), and time delay (without or with time delay) condition.

Discussion

We expected that the effect of behavioral information on the mentally represented facial appearance would be larger in Study 4.1 (with a time delay) than Study 4.2 (without a time delay). The results were only partially in line with this hypothesis. The individual CI ratings showed no difference in effect size between the two studies, whereas the group CI ratings did show weak evidence for the expected difference. Exploratory analyses on the group CI ratings suggested that the hypothesized difference in effect size between the two studies was especially apparent for Faces B and C, but not for Face A.

GENERAL DISCUSSION

We investigated whether behavioral information depicting a person as (un) trustworthy would influence mental representations of that person's face, which was presented for 10 s, more strongly after a time delay than after no time delay. Specifically, we expected that the effect of behavioral information on the mentally represented facial trustworthiness appearance would be notably larger in Study 4.1 (time delay) than in Study 4.2 (no time delay; main hypothesis), present in Study 4.1, and absent in Study 4.2.

Surprisingly, we found effects of behavioral information in both studies with and without time delay. Moreover, the effect was approximately equally large with and without time delay on the individual CI ratings, which contradicts our main hypothesis. The group CI ratings did provide weak evidence in line with our main hypothesis, showing the effect to be larger after a time delay than after no time delay. Explorative analyses including the effect of face identity showed that this predicted interaction effect on the group CI ratings occurred for Faces B and C, but not for Face A. In sum, the data of the individual CIs provide evidence against our main hypothesis, whereas the data of the group CIs are predominantly in line with this hypothesis.

Individual vs. Group CIs

Although the interaction between behavioral information and time delay is statistically speaking not convincing on the individual CI ratings, the pattern of results does look similar for individual and group CI ratings (Figures 4.10 and 4.11). This makes it seem likely that the findings on the group CIs are meaningful. However, recent research points out that the approach of rating group CIs inflates Type I error rates, which does not occur for the approach of ratings individual CIs (Jeremy Cone et al., 2020). Therefore, we advise to interpret the findings on the group CIs with caution and put more weight on the individual CI findings.

As individual CIs are less prone to overestimation of effect sizes than group CIs (Jeremy Cone et al., 2020), it is noteworthy that the main effect of behavioral information was strong enough to emerge on both individual and group CIs in both studies. The biasing effect of behavioral information on mentally represented faces is thus convincing and persistent, at least over a period of approximately 2 days. Although persistent, it does not appear to increase as much as we expected beforehand. The interaction effect did not emerge on the individual CIs, suggesting that the interaction effect must be either smaller than we thought or non-existent.

Effect or No Effect of Behavioral Information after a Face Presentation of 10 s?

Based on the findings of Study 3.2, we started out the present research with the assumption that people mentally represent a person's face quite accurately when the face was presented to them for 10 s. Consequently, we predicted a null effect for Study 4.2. However, although both Study 4.2 and 3.2 presented F. Taylor's face for 10 s and measured the mental representation without any time delay, behavioral information did influence mental representations in Study 4.2. Importantly, this challenges the assumption that viewing a face for 10 s results in unbiased mental representations of that face. We discuss several differences between the two studies to attempt to explain the discrepancy in findings.

First, Study 4.2 counted 300 participants compared to 170 participants in Study 3.2. Perhaps a larger sample in Study 3.2 would have resulted in more evidence for an effect or a smaller sample in Study 4.2 may have resulted in evidence against the effect. Looking at the sequential analysis for Study 4.2, at 170 participants the data provide only moderate evidence for the effect (Figure 4.8). However, the data do show more evidence for the effect under nearly all sample sizes, whereas those of Study 3.2 consistently show more evidence against the effect. Hence, with the current knowledge, it seems that the sample size cannot fully explain the discrepancy in results between the two studies.

Second, Study 4.2 used two other face identities next to the face identity used in Chapter 3. Exploratory analyses on both individual and group CIs show that the effect of behavioral information was approximately equally large for each face identity. This suggests that the addition of Faces A and B does not explain why we found an effect of behavioral information in Study 4.2 but not in Study 3.2. It may, however, explain why we did not find an interaction effect between behavioral information and time delay across Study 4.1 and 4.2. We elaborate on this later in the discussion. Interestingly, the addition of the other two face identities shows that the effect of behavioral information on the mentally represented face generalizes to other face identities than the one used in Chapter 3.

Third, participants in Study 3.2 were instructed to take a good look at the target person's face because they needed to compare it to 500 sets of faces afterwards.

In Study 4.2, participants received no specific instructions before viewing the face. They were simply informed that they would get to see the target person's face for 10 seconds. Consequently, sensory input from the actual face may have been more precise in Study 3.2. Perhaps the effect of behavioral information may have been absent in Study 4.2 as well if we had included a memory instruction, or possibly even better an impression formation instruction (Fiedler, Kaczor, Haarmann, Stegmüller, & Maloney, 2009; Hamilton, Katz, & Leirer, 1980), before the face presentation. Yet, Study 3.3 also lacked a memory instruction and even shortened face presentation time to 100 ms but still lacked convincing evidence for the effect of behavioral information. This suggests that the type of instruction alone would be insufficient to explain the emergence of the effect in Study 4.2.

Fourth, in Study 4.2, participants were instructed to think back to their impression of the person before starting the RC task, likely reactivating their global person or trustworthiness evaluation. The trustworthiness evaluation, which is influenced by the behavioral information as shown in the manipulation check, should be accessible in order to be able to influence the mental representation of the face (Houston & Fazio, 1989). Perhaps the trustworthiness evaluation as manipulated in the current studies is not accessible enough to be activated automatically and therefore only influences the mentally represented face when participants first remind themselves of the evaluation. This could explain why the effect emerges in Study 4.2 including the reminder and not in Study 3.2 and 3.3.

If this reasoning is correct, the effect may occur for newly formed, less rehearsed impressions only if one is reminded of the person impression. Such a reminder is common before making a consequential judgment about a person, such as whether or not someone should get a job, loan, house, or prison sentence. Moreover, it is likely that the participants in Study 3.4 also activated their person impression to visualize a meaningful face prior, which may explain why the effect emerged in that study. Likewise, participants in the study conducted by Hassin and Trope (2000) rated the target person on several personality scales before rating his facial appearance, which arguably activated the person impression as well. Interestingly, the effect of behavioral information on the mentally represented face may occur without the extra reminder of the

impression if the impression is well rehearsed and therefore more accessible, like an established stereotype. Future studies could test this by manipulating the accessibility of the trustworthiness evaluation (activated automatically / effortfully) and the explicit instruction to remind oneself of it (present / absent). Moreover, future studies could manipulate both the type of instruction (e.g. no / memory / impression formation instruction) and the extra reactivation of the impression (present / absent) independently from each other to determine the relative influence on the emergence of the effect.

Alternatively, the explicit instruction, placed shortly before the RC task, may have caused participants to feel that their impression of the target person should be important to their mental representation of the face, creating a demand effect. However, participants had already seen the actual face at this point and the instruction does not mention to what extent the face or behavioral information should inform their impression. Yet, the valenced behavioral information likely influences the impression more than the neutral facial appearance, as evident in the manipulation check. If the instruction indeed created a demand effect, this raises the question to what extent the effect would occur in real life. Perhaps in future research participants could also remind themselves of the actual face before starting the RC task, to eliminate a potential suggestion that they are expected to bias their mental representation of the face in line with their trustworthiness evaluation of the person.

In sum, the most plausible explanation for why we found an effect of behavioral information in Study 4.2 but not in Study 3.2 seems to be the extra reactivation of the trustworthiness impression in Study 4.2, possibly in combination with the lack of a (memory) instruction when viewing the face. The lack of this instruction possibly created a situation of less precise sensory input from the actual face. Assuming that participants use whatever relevant information is available to inform their mental representation of the face, their reactivated trustworthiness evaluation of the person together with whatever they remembered from the actual face could then be used to inform their mental representation of the face.

Why Time was No Convincing Moderator

As noted above, the absence of the interaction effect on the individual CIs suggests that the effect must be either smaller than we thought or non-existent. Several explanations for this unexpected result are conceivable.

Too much weight on person impression

First, the instruction to form an impression about the person during presentation of the behavioral information, the lack of a specific instruction during face presentation, and the instruction to remember one's impression about the person may have put too much weight on this impression opposed to the actual facial appearance. As a result, the effect of behavioral information on the mentally represented facial appearance was immediately large without any time delay, leaving little room for the effect to become even larger over the course of two days. Perhaps a memory (or impression formation) instruction before face presentation would have led participants to encode the facial details much better, resulting in good memory for facial details and hence no effect of behavioral information in the study without time delay. Consequently, there may have been more room for the effect of behavioral information to emerge and grow over time as memory for facial details starts to decline, allowing time to emerge as moderator of the effect of behavioral information on the mentally represented facial appearance.

Selected time delay too short

Second, it is possible that the passing of time does increase the effect of behavioral information on the mentally represented face, but that the chosen time delay of two days was insufficiently long. Perhaps memory for facial details does not deteriorate enough during at least the first two days for the effect to increase notably. It is conceivable that the predicted interaction effect emerges more clearly simply with a longer time delay. Future studies could implement multiple time delay conditions from no time delay at all to delays of days, weeks, and months to see whether time would emerge as a moderator.

Better memory for (un)trustworthy looking faces

Third, it is possible that the predicted interaction effect does not occur for every type of face. We hypothesized that the behavior-based trustworthiness impression would be remembered longer than details about the face (or

behaviors). Hence, when trying to remember what the face looked like after memory for facial details has weakened, the behavior-based trustworthiness evaluation may influence the mentally represented facial appearance. This makes sense for neutrally appearing faces. However, similar reasoning would imply that the interaction effect should be weaker or non-existent for (un) trustworthy appearing faces.

In that case, the face should trigger a face-based trustworthiness evaluation, which too should be easier to remember over time than specific facial and behavioral details. When trying to remember what the face looked like after some time has passed, one may not remember many facial details but simply remember that the face looked (un)trustworthy. Moreover, a recent study suggests that (un)trustworthy appearing faces are recognized better than neutral appearing faces (Mattarozzi, Todorov, & Codispoti, 2015). Consequently, the effect of behavioral information on the mentally represented facial trustworthiness appearance is not expected to increase over time for faces already perceived as (un)trustworthy.

Indeed, Face A, the face identity for which the predicted interaction effect did not emerge, scored relatively high on trustworthiness appearance before the start of the experiment. Moreover, it was the only face identity that appeared to be smiling a little (Figure 4.2), which is associated with trustworthiness appearance (Oosterhof & Todorov, 2008; Todorov & Oosterhof, 2011). The absence of any (interaction) effects of face identity on the manipulation check suggests that the manipulation of the behavior-based trustworthiness impression was equally effective across face identities. Yet, participants may have remembered Face A's facial trustworthiness appearance preventing the mental representation of Face A from appearing very untrustworthy. Indeed, the individual and group CI trustworthiness ratings for face A in the *un*trustworthy condition are similar to or even higher than those for Face B and C in the trustworthy condition.

Alternatively, the predicted interaction effect does theoretically generalize to (un)trustworthy faces, but we did not find it for Face A due to technical issues with the RC task. It is hard to represent a smiling face in a negative way in a RC task. Perhaps participants mentally represented Face A as having a nasty smile in the untrustworthy condition, which was still interpreted as a trustworthy

signal by the raters. Consequently, raters rated Face A's CIs in the untrustworthy condition higher than participants intended. The limited room for bias in the untrustworthy direction may have been enough for a main effect of behavioral information to establish, but not for an additional interaction effect to occur.

In sum, the results for Face A suggest that the interaction effect under investigation may not generalize to all face identities. We emphasize however, given the explorative nature of the face identity analyses, that the results discussed above should be interpreted with caution. Future studies could investigate whether the original trustworthiness appearance of the actual face moderates the interaction effect or whether our results for Face A simply constitute a chance finding.

Theory is incorrect

Fourth, it is possible that the theory is incorrect. We theorized that the behavioral information would become likelier to influence the mentally represented face over time, because memory for facial details would deteriorate quicker over time than the behavior-based trustworthiness impression. Although we did not measure memory for facial details, our manipulation check data suggest that trustworthiness evaluations become less extreme over time as well. If memory for facial details does not decrease much faster than memory for a person's trustworthiness, time delays are unlikely to increase the effect.

Conclusion

Together, the two studies suggest that behavioral information influences mental representations of faces to appear more in line with this information, even immediately after 10 s of face presentation. That is, if one focuses on forming an impression of the person and reminds oneself of this impression, which people usually do when they have to make an important decision about a person. Interestingly, the effect of behavioral information on the mental representation of a face that was presented for 10 s is thus more prevalent than we expected on the basis of our earlier work (Chapter 3). This suggests that, when making a decision about someone who supposedly behaved (un)trustworthily, that person's facial appearance looks slightly more (un)trustworthy in one's mind than the actual facial appearance of that person.

Furthermore, the studies carefully hint that this bias *may* become stronger after more time passes since seeing the face, at least for neutral-appearing faces. This is in line with our hypothesized theory that memory deteriorates faster for details than global person evaluations. However, given the mixed and weak confirmatory evidence, as well as the exploratory nature of the face identity findings, more research is needed to provide clearer evidence for or against this hypothesis.

If the suggestions above are supported by future research, people can take this into account when making decisions about others. For instance, the results suggest that seeking out reliable information to form impressions about a person's trustworthiness may help alleviate the unwanted impact of inaccurate face-based impressions by bringing the mentally represented facial appearance a bit more in line with the person's trustworthiness. At the same time, the findings provide a warning for the influence of inaccurate/misinterpreted behavioral information and for instances in which accurate mental representations of the face are crucial, as with eyewitness reports.

APPENDIX 4A

Behavioral descriptions are taken and translated from Dotsch et al. (2013) and Fuhrman, Bodenhausen, and Lichtenstein (1989). One description was offered by colleagues from our lab group. Descriptions marked with (U) are intended to appear untrustworthy, (N) to appear neutral, and (T) trustworthy.

A pilot study validated that the untrustworthy set (M = -70.91, SD = 39.58) was indeed evaluated as more untrustworthy than the trustworthy set (M = 56.09, SD = 24.97) on a scale ranging from -100 (*untrustworthy*) to 100 (*trustworthy*): δ = -2.54 [-4.03; -1.25], BF_{-0} = 13380.74.

Untrustworthy:

1. Threatened another person with a knife (U)
2. Robbed another person in an alley (U)
3. Provoked a fight while going out (U)
4. Stole another person's jacket at a party (U)
5. Swore at the bus driver (U)
6. Pulled the seat out from underneath somebody (U)
7. Smoked in a non-smoking section even though others complained (U)
8. Tripped someone in the hall (U)
9. Ate a sandwich at the station (N)
10. Crossed the street (N)

Slightly untrustworthy:

1. Shoved a man who was passing out leaflets (U)
2. Pushed another person out of the way to walk through (U)
3. Sneaked in front of people in the queue at the checkout (U)
4. Drank a glass of water (N)
5. Waited at the bus stop (N)
6. Opened the window a little (N)
7. Went grocery shopping at the mall (N)
8. Cleaned the dishes (N)
9. Sent a text message to a friend (N)
10. Had a chat with a neighbor (N)

Slightly trustworthy:

1. Called the neighbor to say that someone tried to break into his car (T)
2. Never smoked in other people's apartments (T)
3. Gave money to charity (T)
4. Took a stroll down the dock (N)
5. Took a can of Coke from the vending machine (N)
6. Locked the door (N)
7. Looked for the bike in the parking (N)
8. Replaced the kitchen light bulb (N)
9. Emptied the trash (N)
10. Turned out the lights before going to bed (N)

Trustworthy:

1. Returned a found wallet (T)
2. Warned a woman that she dropped 50 dollar (T)
3. Volunteered to collect money for poor children (T)
4. Supported a friend in difficult times (T)
5. Lived up to promises made, even if this took a lot of effort (T)
6. Told the cashier that she gave too much change (T)
7. Acted politely and kind to a stranger in the street (T)
8. Helped a man picking up the groceries the man dropped (T)
9. Got an umbrella out as it started to rain (N)
10. Stepped into the elevator (N)

APPENDIX 4B

In the manipulation check, participants were asked to provide their impression of the target person (F. Taylor) on the following scales:

- Bad – Good
- Incompetent – Competent
- Submissive – Dominant
- Untrustworthy – Trustworthy (scale of interest)
- Unintelligent – Intelligent
- Not criminal – Criminal
- Cold – Warm

The scales ranged from -4 to 4.


5

General Discussion

In the present dissertation, we aimed to increase understanding of the circumstances in which people 'read into faces' (RIF) based on verbal information about the face bearer. Specifically, we aimed to increase understanding of *the circumstances under which verbal information about a person's behavior is more vs. less likely to bias one's visual mental representation of that person's seen face*. We approached this main research question with a Bayesian inspired theoretical view and attempted to visualize approximations of participants' mental representations using a data-driven reverse correlation (RC) methodology.

The RC methodology visualizes approximations of participants' mental representations in so-called classification images (CIs), which need to be rated by an independent group of raters. Unfortunately, this rating method is cost-inefficient when conducting sequential hypothesis testing. Before investigating the main research question, we therefore introduced and tested an innovative, cost-efficient alternative to this rating method, called the criterion creation method, in Chapter 2. We demonstrated how to create and validate a criterion on which CIs can be scored and compared its results to the rating method. Although the criterion creation was more efficient than the rating method when conducting sequential hypothesis testing, the criterion creation method was also arguably less sensitive. We proposed a combination of both methods for the most optimal and efficient test, allowing researchers to profit from the use of both the RC task and sequential hypothesis testing. After this methodological contribution, the remaining chapters focused on the theoretical contribution of this dissertation by investigating the main research question.

With regard to the main research question, the Bayesian inspired view generated two successive general predictions, namely the 'Face Prior Hypothesis' and the 'Prior-Likelihood Balance Hypothesis'. The Face Prior hypothesis stated that if verbal information is to influence the visual mental representation of the seen face at all, the verbal information should generate a visual expectation of the facial appearance, also called a face prior. Subsequently, the Prior-Likelihood Balance Hypothesis stated that the verbal information is more (less) likely to bias the mentally represented facial appearance of the seen face if this face prior is relatively strong (weak) compared to input from the seen face. We investigated the two general predictions over the course of six studies in Chapters 3 and 4.

In this chapter, we will review what we have learned from the present dissertation regarding the two general predictions and consequently regarding the main research question. Do the data provide evidence for or against the general predictions and which specific circumstances appear to make the bias (dis)appear? Furthermore, we will discuss theoretical and societal implications of the current work as well as emerging questions regarding the process of RIF. Last, we will critically reflect on methodological and analytical choices made in the present research.

EVIDENCE REGARDING THE TWO GENERAL PREDICTIONS

The Face Prior Hypothesis

The Face Prior Hypothesis states that verbal information about a person can trigger a visual expectation about that person's facial appearance, or face prior. Scientific efforts increasingly endorse the idea that humans represent knowledge largely perceptually (Barsalou, 1999, 2008; Seifert, 1997; Shepard & Metzler, 1973; Wheeler, Petersen, & Buckner, 2000), demonstrating how even abstract concepts can be represented in a perceptual system (Barsalou, 1999). Together with empirical findings that humans associate conceptual knowledge about others (e.g. about their personality, race, gender, age, health, and emotion) with specific facial appearances (DeBruine, 2002; Johnson et al., 2012; Sutherland et al., 2013; Todorov et al., 2015; Zebrowitz, 2017), it seems likely that conceptual knowledge about a person can activate a visual expectation of facial appearance.

The Face Prior Hypothesis was tested in Study 3.1. In this study, we manipulated verbal information about a target person's behavior with the aim to generate a trustworthiness impression (untrustworthy / trustworthy) about this person, confirmed in a manipulation check at the end of the study. After the verbal information, we measured participants' expectations of the person's facial appearance (i.e. their face priors) by having them visualize their mental representations of his face in a two-images forced choice RC task (Dotsch & Todorov, 2012) without ever having seen his actual face.

The data from Study 3.1 provided overwhelming evidence for the Face Prior Hypothesis. According to this study, it appears that verbal information about a person's behavior can indeed generate an expectation about that person's facial

appearance (i.e. a face prior). With this finding, the present research supports existent research on the effect of information about group membership on the expected facial appearance (Brown-Iannuzzi et al., 2016; Dotsch et al., 2008, 2013; Kunst et al., 2017; Ratner et al., 2014). Moreover, the present research goes beyond the concept of group membership, showing that information about the behavior of one specific individual can influence the expected facial appearance for that specific individual.

The Face Prior Hypothesis formulated a prerequisite for the effect of verbal information on the mentally represented facial appearance of a seen face to occur. Adopting a Bayesian inspired view helped to make this prerequisite explicit. Next, the Prior-Likelihood Balance Hypothesis is directly concerned with the main research question of the present dissertation, as it focuses on the *circumstances* under which the effect may be expected to (dis)appear.

The Prior-Likelihood Balance Hypothesis

The Prior-Likelihood Balance Hypothesis is based on an idea consistent with Bayesian theories (Clark, 2013; Mamassian et al., 2002) that is also advocated in a Dynamic Interactive (DI) theory of social perception (Freeman et al., 2020). The idea is that the mental representation of a seen face is influenced both by bottom-up sensory input from the actual face that is 'out there' (the likelihood) and by the perceiver's top-down expectations about the facial appearance (the prior). The Prior-Likelihood Balance Hypothesis states that the relative strength of these two components determines how much the prior expectations will bias the mental representation of the seen face.

The present research tested this hypothesis with regard to the influence of verbal information about a target person's behavior on the trustworthiness appearance of the target's face. Imagine a dimension that portrays faces varying on trustworthiness, from untrustworthy to trustworthy looking faces. Based on the verbal information about the target's behavior, participants may expect him to be (un)trustworthy and may expect to see a rather (un)trustworthy looking face. Their prior distributions would consequently peak on the (un)trustworthy end of the facial trustworthiness dimension. The stronger their expectation, the higher and narrower the peak. The target's actual face was selected to look rather neutral on trustworthiness. The likelihood distribution should therefore

peak in the neutral middle of the facial trustworthiness dimension. The stronger the sensory input from the actual face, the higher and narrower this peak. The prior and likelihood together influence the posterior distribution, which determines the final mental representation of the face. Our goal was therefore to manipulate the relative strength of the prior and likelihood and investigate the impact on participants' mental representations of the target's seen face. Figure 5.1 illustrates the idea that the relative strength of the prior and likelihood matters for the posterior distribution.

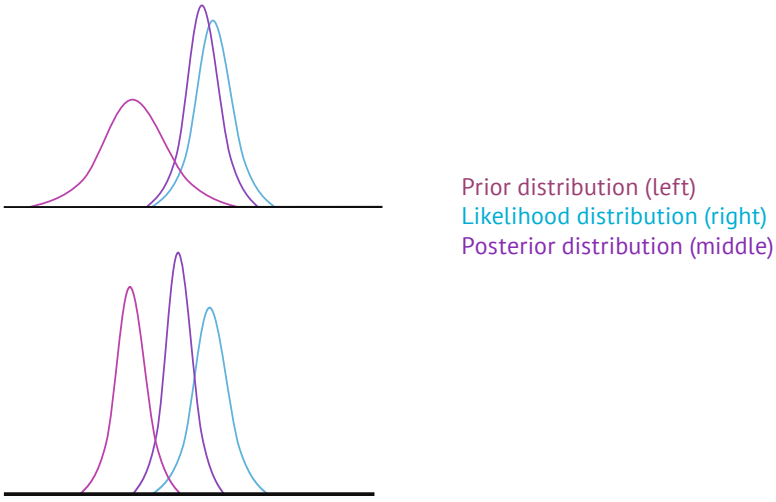


Figure 5.1. Illustration of the idea that the balance in strength of the prior (left distribution in pink) and the likelihood (right distribution in blue) impacts the posterior (middle distribution in purple) distribution. As the prior distribution becomes more precise from the upper to the lower graph, the posterior distribution is increasingly drawn to the prior.

We tested the Prior-Likelihood Balance Hypothesis with varying operationalizations across five studies. All studies had a similar set-up as Study 3.1, except that we presented the actual face before participants started the RC task, so the actual face could inform participants' mental representation of the face. The pattern of results in Chapter 3 was in line with the general idea that verbal information will bias the mentally represented face if the face prior is relatively strong compared to input from the actual face. Study 3.2 to 3.4 showed an increase in evidence for a RIF effect as we attempted to make the face prior relatively stronger (in relation to input from the actual face) with each study. As starting point, Study

3.2 investigated the effect of verbal information on the mental representations of the seen face when participants were encouraged to remember the actual face well (memory instruction) and the actual face was presented for 10 s. Although we expected to find a RIF effect under such circumstances based on the work by Hassin and Trope (2000), we found evidence against it. Attempting to weaken input from the actual face (i.e. weaken the likelihood), Study 3.3 eliminated the memory instruction and reduced face presentation duration to 100 ms. This resulted in mixed evidence for a RIF effect. Attempting to strengthen the face prior, Study 3.4 added an instruction to visualize one's expectation of the face before presenting the actual face for 100 ms. This time, the data provided evidence for a RIF effect.

In Chapter 4, we investigated a different operationalization to weaken the likelihood in comparison to the prior, namely the amount of time passed after both verbal information and the face had been presented but before measuring the mental representation in the RC task. The idea was that the person impression of the target as (un)trustworthy (influenced by the verbal information and probably used to inform the face prior) should remain relatively stable over time, whereas memory for the target's actual facial features should decrease over time. This should result in a stronger RIF effect as more time passes. Although the data carefully suggested that a time delay of approximately 2 days could perhaps lead to a stronger RIF effect, the data remained unconvincing, and more research is needed to clarify this. Instead, the studies seemed to demonstrate the importance of reactivating the person impression for a RIF effect to emerge (to which I will return later). The set-up of Study 4.1 was similar to Study 3.2 in that it presented the face for 10 s, but it eliminated the memory instruction and instructed participants to think back to their impression of the person after both verbal information and the actual face had been presented but before starting the RC task. It also added a time delay of approximately 2 days after viewing the actual face, attempting to weaken the impact of the actual face. As expected, we found evidence for a RIF effect. Study 4.2 was identical to Study 4.1 except that it eliminated the time delay. Although we had expected the 10 s face presentation to this time result in evidence against a RIF effect (as in Study 3.2), we surprisingly again found evidence in favor of a RIF effect under these conditions. Taking the data of both studies together, we did find weak evidence for a stronger effect after the time delay on one measure,

namely on the ratings of the RC group classification images (CIs), which consist of one averaged image per experimental condition. However, we did not find this on our most reliable measure, namely on the ratings of the RC individual CIs, which consist of one averaged image per participant. Importantly, although we did not find convincing evidence that the RIF effect increased over time, the effect did not become weaker either, which would have contradicted the Prior-Likelihood Balance Hypothesis.

On the basis of Chapters 3 and 4, we suggest that the balance between the face prior (informed by the verbal information) and input from the actual face seems to matter for the emergence and strength of RIF. One may notice that the Prior-Likelihood Balance Hypothesis can in principle explain any single research finding in hindsight: "Ah, we did (not) find a RIF effect, so apparently the prior was (not) strong enough in comparison to the likelihood". This may create the impression that the hypothesis is merely descriptive and lacks predictive power (i.e. is unfalsifiable). Importantly though, the hypothesis does make clear predictions about *patterns* of results. If one study is explained in hindsight, then altering only one aspect of this study (e.g. strengthening the prior) should change the balance between the prior and likelihood and therefore change the results in a predictive manner. For instance, strengthening the prior while keeping the likelihood constant should either lead to a similar bias (if the strengthening is not substantial enough) or to a stronger bias, but never to a weaker bias. This is demonstrated nicely in the patterns of results in Chapter 3 and 4. If the effect had grown weaker across studies in Chapter 3 or with time delay in Chapter 4, this would have spoken against the Prior-Likelihood Balance Hypothesis.

Besides being useful for predicting patterns of results, we considered the Prior-Likelihood Balance Hypothesis useful for guiding specific operationalizations of circumstances that make the effect of verbal information on mental representations of a seen face more (less) likely to occur. Moreover, the hypothesis proves helpful in organizing and understanding the research findings. Using the hypothesis as organizing framework, what have we learned from the present research about specific circumstances that make the RIF effect more vs. less likely to occur?

Factors that change the prior strength

To strengthen the prior, it seems essential that participants activate their impression of the person, either about the person in general (Chapter 4) or specifically about the expected facial appearance (Study 3.4). Comparing Study 3.3 and 3.4 suggests that visualizing the expected facial appearance after the behavioral information and before face presentation was critical for the effect to occur. However, comparing Study 3.2 to Study 4.2 suggests that reactivating one's person impression after both the behavioral information and face presentation is already sufficient for the effect to emerge. Whereas it appeared quite difficult to find a convincing RIF effect in Chapter 3, where a RIF effect emerged only when participants mentally visualized their face priors, it appeared quite difficult *not* to find a RIF effect in Chapter 4. As the most striking difference in the set-up of these Chapter's studies is that participants always activated their person impression in Chapter 4, the accessibility of the person impression seems crucial for the RIF effect to appear.

Indeed, inspection of the two publications demonstrating RIF effects mentioned in Chapter 1 (Hassin & Trope, 2000; Levin & Banaji, 2006) shows that the relevant person impressions were explicitly pointed out to participants (and thus activated) shortly before evaluating the face. In the study reported by Hassin and Trope (2000), participants rated the target person on several personality traits, including three kindness related traits, after reading the verbal information describing the target as (un)kind and before rating the facial appearance. This probably increased participants' focus on the target's personality, including his kindness. In the study reported by Levin and Banaji (2006), participants were explicitly told beforehand that the study was about "how people perceive the shading of faces of different races". Additionally, the faces were explicitly labeled 'Black' or 'White'. This likely increased participants' focus on race. It can thus be argued that the person impressions intended to influence perception of faces were highly accessible in these studies.

Future research could further experiment which circumstances lead to stronger (or more accessible) person impressions. If the person impression is strongly associated with facial appearance, this should lead to stronger face priors as well. As suggested in previous chapters, researchers could for instance compare the use of elaborate, vividly described scenarios vs. one-sentence scenarios, or

of well-established stereotypes vs. newly formed impressions. For instance, it can be expected that stereotypes of race, age, and sex are well-established and will often be activated automatically (Quinn et al., 2007). Another factor open to investigation is the order of presentation of the verbal information and the face. Perhaps presenting the verbal information last allows for stronger effects of the information on the mentally represented facial appearance. Yet other factors open to investigation include the perceived diagnosticity of verbal information about behavior for facial appearance, the reliability of the information source (e.g. a known gossiper vs. professional or someone know to resent / love / feel neutral towards the person), extent of elaboration on the verbal information, content of the verbal information (e.g. information implying target's personality, occupation, race, gender, or age), personal relevance (e.g. you later have to interact with or evaluate the person, or the described behavior was directed at you or a loved one), an expectation-(in)congruent context during face perception (e.g. positive/negative background), a goal to (dis)confirm one's expectations, and participants' mood (Jeremy Cone & Ferguson, 2015; Huang & Bargh, 2014; Kunda, 1990; Quinn et al., 2007).

Factors that change the likelihood strength

To weaken the likelihood, it may be efficient to shorten exposure to the actual face (Studies 3.2-3.4), but only sufficiently so when combined with an activation of one's expectation of the person's facial appearance (Study 3.3 vs. 3.4). Moreover, activation of one's impression of the person seems so crucial, that it may not even be necessary to shorten exposure to the actual face (Chapter 4). Indeed, the above referenced earlier work on RIF, in which the relevant person impression was likely activated, showed that the effect occurred even under unconstrained perception of the face (Hassin & Trope, 2000; Levin & Banaji, 2006). A time delay of approximately 2 days after face presentation does not convincingly weaken information on the actual face any further in relation to the prior, at least when activation of the person impression presumably results in an already relatively strong prior (Chapter 4). To strengthen the likelihood, it might help to implement a goal to accurately remember the facial appearance (Study 3.2), but we have not investigated this in combination with a reactivation of one's impression of the person. It would be interesting to see how those two manipulations combine.

Future research could investigate whether a longer time delay does weaken information from the face relatively strongly or whether a 2-day time delay is sufficiently long when participants do not activate their impression of the person. Future research could also investigate the effect of varying face presentation durations and of a memory instruction (yes / no) when combined with a reactivation of the person impression. As suggested in previous chapters, other potential factors open to investigation are less optimal viewing conditions of the face that are more representative of real-life situations (e.g. shadows on the face, objects obstructing part of the face, non-frontal view of the face, head movements, distractions in the environment), and the extent to which participants feel their memory of the face is tested on accuracy (which probably is not the case in most real-life situations, but appeared to be the case for participants in the RC task). Figure 5.2 depicts the schematic overview of the general predictions from Chapter 1 with examples of manipulations and measurements from our studies added in red, giving an overview of what we have learned so far.

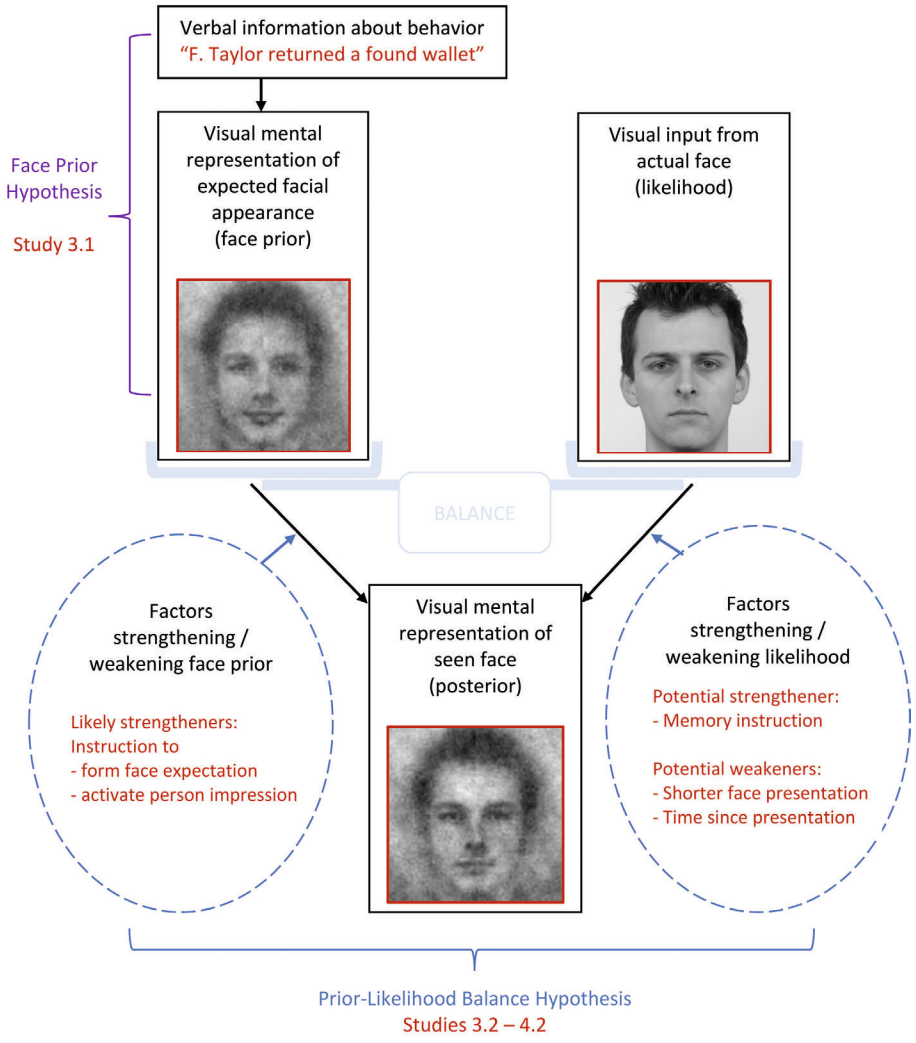


Figure 5.2. Schematic representation of the two general predictions taken from Chapter 1 with examples of manipulations and measurements from our studies added in red (likelihood and posterior examples taken from Study 3.4). Due to space restrictions, only examples from the 'trustworthy' experimental condition are portrayed. The Face Prior Hypothesis was tested in Study 3.1. The Prior-Likelihood Balance Hypothesis was tested in Studies 3.2-3.4 and 4.1-4.2. The varying circumstances which were aimed at manipulating the relative strength of the expected facial appearance (face prior) and input from the actual face (likelihood) are added in the blue circles with indications of their successfulness.

Intermediate conclusion

In conclusion, it appears that the face prior should be strong enough compared to input from the actual face for verbal information to bias the mental

representation of a seen face. For this to be the case, the most important factor seems to be that the person impression (informed by the verbal information) should be accessible when mentally (re)constructing the seen face. Possibly, one also should not have the goal to remember the face very accurately, but further research is needed to investigate the effect of a memory goal in combination with an accessible person impression.

The importance of the accessibility of the person impression is in line with literature on attitude and category (or stereotype) accessibility (or activation), showing that the attitude or category should be accessible for it to influence evaluations (Bruner, 1957; Houston & Fazio, 1989; Quinn et al., 2007). It also agrees with recent research showing that the biasing impact of stereotypes on visual representations of faces depends on the functional interaction between brain areas visually representing faces (e.g. fusiform gyrus) and brain areas holding the conceptual stereotypes (medial orbitofrontal cortex; Barnett, Brooks, & Freeman, 2020). In other words, the bias in the face representation seems to depend on the accessibility of the conceptual stereotype. The concept of accessibility teaches us that even if the verbal information generates a person impression that is associated with specific facial features, this will bias the mentally represented seen face only if the impression is accessible (i.e. activated) at the time of mentally representing the seen facial appearance. Of course, the stronger the impression, the more likely it will be activated automatically and hence will be accessible (Fazio, Sanbonmatsu, Powell, & Kardes, 1986).

The present research used rather extreme single sentence behavioral descriptions to generate person impressions. Still, it is questionable how strong resulting impressions are compared to for instance an elaborate, vividly described scenario or to established person impressions that have evolved over many instances in life. If the latter methods lead to more accessible impressions, they may bias the mentally represented seen facial appearance even without an explicit reminder of the impression. Newly formed, less rehearsed person impressions, on the other hand, may bias mentally represented seen faces only if one reminds oneself of the impression. Of course, people often may do this before making a consequential decision about a person, such as hiring, financial, and juridical decisions.

THEORETICAL AND SOCIETAL RELEVANCE

Besides providing support for the general predictions derived from the Bayesian inspired view, the present research supports and extends existent literature in at least four manners. Moreover, these theoretical insights suggest several societal implications.

Humans as Active Perceivers

Firstly, it is noteworthy that some studies (Study 3.4, 4.1, and 4.2) demonstrated a convincing effect of verbal information on the mental representation of a seen face. It appears that conceptual information can indeed be read into faces. Interestingly, the biases in these visualized mental representations were much smaller than when participants had not seen the actual face and thus visualized their expectation of an unseen face (Study 3.1). After seeing the actual face, participants seemed to incorporate information from the actual face into their mental representations, retaining a subtle bias only. These findings are in line with the Dynamic Interactive (DI) theory that social cognitive processes in the perceiver's mind and visual input together influence social person perception (Freeman et al., 2020). Moreover, they correspond with the broader notion that human perception is an active reconstruction of what is 'out there' (e.g. Allport, 1954; Bruner, 1957; Clark, 2013; Friston, 2010; Hohwy, Roepstorff, & Friston, 2008; Lippmann, 1922; Mamassian, Landy, & Maloney, 2002).

The literature above suggests that RIF occurs during perception. As we measured mental representations afterwards in the RC task, not live during perception of the face, we cannot know whether the effects in our studies occurred during perception or afterwards during mental reconstruction. Importantly though, thanks to our experimental set-up we do know that the bias probably exists even after perception of the actual face, which is usually when people make influential decisions about others, such as whether or not to vote for or hire someone, or how to describe someone's face in eyewitness procedures. Although we have not measured such decisions in our studies, existent research suggests that the mentally represented facial appearance can influence these decisions (Antonakis & Eubanks, 2017; Hassin & Trope, 2000; Todorov et al., 2015). Consequently, RIF may have impacting social consequences irrespective of whether it occurs during perception or shortly afterwards in memory. Moreover, the findings of

Chapter 4 teach us that if the bias did appear during perception already, it is not one that emerges momentarily only to be corrected quickly. Instead, it is a bias that can persist for at least 2 days.

For society, the present findings highlight that our experience of the world is a potentially colored one. Apparently, besides "seeing is believing", sometimes "believing is seeing" as well. Hopefully, awareness that human perception is an active reconstruction may foster compassion for others who advocate different opinions, decisions, and behaviors. After all, our prior experiences shape our mental models, which shape our current experiences in life. Consequently, if I had lived the life of the person whose behavior I currently condemn, perhaps I would have behaved similarly, because with that person's worldview, it may seem the most suitable thing to do. This realization may help me to refrain from judgment and instead try to look past the behavior, to understand its drive and intention. In short, by accepting that our perceptions can be colored we may be able to increase understanding and compassion for others that differ from us.

Faces as Both Influential and Influenced Social Objects

Secondly, a vast literature exists on the phenomenon 'reading from faces' (RFF; Hassin & Trope, 2000): the impact of facial appearance on social impressions, decisions, and behaviors (Sutherland et al., 2013; Todorov, 2017; Todorov et al., 2015). The present dissertation extends this literature by demonstrating that faces in themselves can be influenced by social information. Importantly, this was the case even for an individual's face that was actually seen, supporting the small number of existent studies on RIF (Hassin & Trope, 2000; Levin & Banaji, 2006). What a face objectively looks like is thus only part of the story in social face perception. This is in line with recent research demonstrating that face based impressions depend not only on target characteristics, but for a large part also on perceiver characteristics and target x perceiver interactions (Hehman et al., 2019).

From a societal point of view, this can be interpreted both as good and bad news. As good news, the findings suggest that people are not totally helpless with regard to the (dis)advantageous effects of their natural facial appearances. Information about a target (e.g. suggesting trustworthiness) may subtly bias perceiver's mental representation of the target's face in line with the provided

information. Given that the provided information may be more representative than the target's natural facial appearance of the target's actual personality, intentions, and behaviors, this biasing effect is beneficial for both target and perceiver. The target suffers a little less from inaccurate evaluations based on his or her natural facial appearance and the perceiver has a somewhat more representative basis for evaluating the target. Future research should extend our findings, which were demonstrated on neutral faces, to faces that naturally appear more pronounced on the dimension of interest (e.g. untrustworthy).

If the verbal information is in fact unrepresentative of the target, however, the biasing effect is of course undesirable. The present findings suggest that such information may get engrained in the visual representation of the face, which may strengthen the information-based impression. Perhaps it could make the impression even more resistant to change, although this is debatable (Barnett et al., 2020). On the other hand, the present findings also suggest that biases in the mental representation of the face are subtler once perceivers saw the individual's face compared to when they never saw the face. Now that we have shown it is likely that RIF biases exist (under certain circumstances), future research can estimate their magnitude more precisely and investigate to what extent such subtle biases really lead to social consequences.

Irrespective of whether the information is representative of the target or not, the bias is always bad news when an accurate representation of the face is crucial, as with eyewitness procedures. It is highly important to accurately represent the face for increasing chances of catching the perpetrator. Unfortunately, it is likely that the mental representation of the perpetrator's face gets biased by the impactful negative behavior of the crime as this kind of behavior probably creates a strong and accessible person impression about the perpetrator. If so, eyewitness' descriptions of perpetrators' faces are likely to be biased. An interesting question for future research would be whether or not such biased mental representations may also affect perceivers' (rate of) face recognition on a later encounter with the same person.

Verbal vs. Visual Information

Thirdly, the present research findings support existent research by showing that *verbal* information can influence mental representations of a seen face

(Hassin & Trope, 2000; Levin & Banaji, 2006), just like visual information can (Bijlstra et al., 2014, 2010; Freeman et al., 2011; Johnson et al., 2012; Van den Stock et al., 2007). This finding indicates that mental representations of faces can be influenced by person-relevant information that was not even present in the visual field during face perception. This strengthens the theoretical notion that our visual experience of the social world is an integration of both visual input and conceptual knowledge (Freeman et al., 2020). Verbal information about others is highly prevalent in our everyday social lives and can be spread by anyone (not just the target person). Moreover, verbal information about a target can easily be made up while it is much harder for the receiver to check the information's truthfulness. It thus appears based on our findings that there exists a source of influence on how people mentally represent others' faces that is prevalent, hard to constrain, and easily manipulated.

RIF as Context-Dependent Process

Fourthly and importantly, the present research goes beyond existent research by showing that the effect of RIF is *context-dependent*, (dis)appearing according to a broadly predictable pattern. As such, the present research shows the merit of adopting a Bayesian inspired view in social face perception research. Apparently, there are situations in which people's mental representations of a seen face are more accurately than we previously expected on the basis of earlier RIF findings (Hassin & Trope, 2000; Levin & Banaji, 2006). At the same time, there are situations in which these mental representations are more biased than would be expected based on the social face perception literature, which often manipulates facial appearances to investigate their impact on social impressions and consequences (Todorov, 2017). With the present research, we have made a start in uncovering the specific circumstances that make the RIF effect more vs. less likely to occur, showing that the (in)accessibility of the person impression at the time of mentally representing the face plays an especially important role.

Given that RIF may have significant social consequences, it is relevant to understand the circumstances under which RIF can be expected to (dis)appear. Subsequently, we can try to incorporate this knowledge into social decision-making processes. For example, when someone's facial appearance should not matter for the decision (which should be the case for most social decisions in society), our findings highlight the importance of seeking out reliable and

relevant information to inform one's impression of the person (e.g. of the person's trustworthiness). Moreover, it may even be advisable to focus on this reliable information before viewing the face. This way, a potential RIF bias may even be helpful as it may bring the mental representation a bit more in line with the person's actual trustworthiness. At the same time, our findings highlight caution for information that may be unreliable or easily misinterpreted as an inaccurate bias in the mental representation is clearly unwanted.

When someone's facial appearance is highly relevant to the decision, as is the case for eyewitness reports, our findings provide a clear warning for the potential biasing effect of information about the person. Returning to the hypothetical example of the robbery from Chapter 1, what can we say about your mental representations of the perpetrator's face based on the present dissertation? Under which circumstances may your mental representation of the face (not) be influenced by the information about the robbery? Assuming that you believed that someone indeed was robbing a store, your expectation of the perpetrator's face probably looked somewhat untrustworthy. Next, the extent to which this expectation biased your mental representation after seeing the perpetrator's face depends on the strength of this expectation as well as how well you got to see the perpetrator's face. If your expectation of the face was not active when mentally representing the perpetrator's seen face, your mental representation would probably be quite accurate. On the other hand, if the perpetrator's behavior was still on your mind and led you to strongly expect the perpetrator to look untrustworthy, chances are high that the perpetrator's face ended up slightly biased in your mind, even if you got a good look at the actual face. Of course, you probably also had a strong goal to accurately remember the perpetrator's facial appearance. It is unclear from the present dissertation if you would also mentally represent the face biased under these circumstances.

Based on our findings, we would advise police officers interviewing eyewitnesses to first get the face description and encourage eyewitnesses to focus on the face alone. It may be unwise to describe the perpetrator's behavior or otherwise activate the eyewitness' person impression of the perpetrator before getting the face description. Unfortunately, a perpetrator's behavior is often impactful and therefore likely to be accessible in the eyewitness' mind when viewing and describing the facial appearance. Moreover, eyewitnesses scarcely have the

chance to look at the perpetrator's face under such optimal viewing conditions as in the present studies. Our findings suggest that police officers should therefore be aware of biases in the eyewitness' reports. Relatedly, once a suspect is arrested, police officers should be aware that if the lineup contains fillers who look less criminal than the suspect, eyewitnesses may identify the suspect as the perpetrator simply because the suspect has the most criminal looking face (Flowe & Humphries, 2011; Wagenaar, 1989).

EMERGING QUESTIONS ON THE PROCESS OF RIF

The present research provided some answers and suggestions regarding the circumstances under which RIF is more and less likely to occur. Like most research, it gave rise to many questions too. We introduce some theoretical questions regarding the nature of the RIF process that arose in response to the present research.

Prior Accessibility or Demand Effect?

As discussed above, the accessibility of the person impression appeared as an important moderator for RIF to occur. However, the accessibility of the person impression was always manipulated through an explicit instruction in the present research. Therefore, it is conceivable that the emergence of RIF in Study 3.4 and in both studies of Chapter 4 is each time due to a demand effect. If so, the effect is unlikely to occur in real life situations. Although not impossible, inspection of the questionnaire data renders the explanation of a demand effect unlikely. Participants seemed highly focused on remembering the target person's face accurately, which is apparent from their responses in which they indicated their focus on specific facial features they remembered him to have and their concerns that it became harder for them to accurately remember what he looked like as the RC task progressed. Some participants even suggested to next time "please show his face at the end", so they could see whether they had remembered the face accurately. These responses suggest that most participants tried their best to remember the actual face accurately, not change it in line with the behavioral information.

To exclude the possibility of a demand effect, future research could use verbal information that triggers a person impression that is highly established and

therefore chronically accessible and thus automatically activated upon thinking about or encountering the person. Using a chronically accessible impression may eliminate the need to explicitly instruct participants to think about their impression of the person and thereby minimize the probability of a demand effect. Strong stereotypes are good candidates for highly accessible impressions. For example, the target person could be described in a vignette as a violent hooligan or caring nurse. Alternatively, as mentioned in Chapter 4, future research could instruct participants to not only think back of their impression of the target person, but also to think back of what the target person's face looked like, before starting the RC task. This addition might minimize a potential feeling on the side of participants that they should bias their mental representation of the face in line with their person impression.

Other Potential Boundary Conditions

Chapters 3 and 4 investigated several boundary conditions on the (dis) appearance of the RIF effect, namely the influence of the accessibility of the person impression (informed by the verbal information), a goal to accurately remember the face, the face presentation duration, and passage of time since face presentation. Besides these conditions, the chapters generated new insights into other potential boundary conditions.

Extent of (dis)similarity between expected and actual facial appearance

The present research always found an assimilation effect if an effect emerged. Assimilation means that the mentally represented facial appearance was biased in the direction of the information describing the person as (un)trustworthy, not in the opposite direction, which is called a contrast effect. Assimilation effects are often found when the expected and actual facial appearance are somewhat similar, not too different from each other (e.g. Freeman, Penner, Saperstein, Scheutz, & Ambady, 2011; Johnson, Freeman, & Pauker, 2012; Levin & Banaji, 2006). We used a relatively neutral face as actual face, which arguably is ambivalent enough to be interpreted in both directions.

Indeed, we noticed some instances in the questionnaire data where the eyes and mouth were actively interpreted in line with the experimental condition. For instance, participants in the trustworthy condition would mention that the target person, who was always called F. Taylor, was smiling friendly, whereas

participants in the untrustworthy condition mentioned him to be smirking or not smiling at all. Similarly, depending on the experimental conditions, F. Taylor was often described as having "stern", "hard", or "cold eyes", or having a "harsh" or "hard look in the eyes" (untrustworthy condition), opposed to having "kind eyes" or a "friendly expression" (trustworthy condition). For these participants, the neutral face indeed appeared ambivalent enough to be interpreted in line with the verbal information they had received.

On the other hand, the questionnaire data showed that for some participants, the actual face appeared different from the suggestion based on the verbal information. For instance, some participants in the untrustworthy condition mentioned that F. Taylor did not look "mean" or "bad", did not have a "hard sulky evil face", or even remembered him to have "a friendly looking face". Some participants in the trustworthy condition mentioned that F. Taylor looked "quite serious and stern", "almost angry", "sad", "too grumpy", "not happy", "not particularly warm", and "surprisingly kind of hard edged, considering his very decent behavior", remembering "the hardness of his face despite being a nice person".

If some participants noticed these differences between their expectation and the actual face when the actual face looked rather neutral, what would happen if the actual face looks extremely (un)trustworthy whereas the verbal information suggests the opposite? Interestingly, Hassin and Trope (2000) still found assimilation effects even for unambiguous faces. Indeed, theory on social comparison processes suggests that in most social comparison situations, the default is to focus on similarities, resulting in assimilation effects (Mussweiler, 2003). Yet, the same theory suggests that if the standard (in this case the expected facial appearance) and the target (in this case the actual face) differ remarkably, especially the differences between the two should stand out, resulting in a contrast effect (Mussweiler, 2003). Moreover, research on conceptual stereotypes suggests that activation of the person impression may not only facilitate processing of impression-congruent information (in this case impression-congruent facial features), but may also help to allocate attention to impression-incongruent information (Quinn et al., 2007). The activated person impression could thus help in processing the unexpected facial features.

As yet, it thus remains unclear what kind of effect is to be expected when the expected and actual facial appearance differ remarkably. Do our cognitive systems use the bottom-up sensory input from the actual face to correct the expected facial appearance (resulting in a veracious mental representation)? Or do our cognitive systems focus on the differences resulting in a contrast effect? Or is processing of impression-congruent facial features still stronger than that of incongruent ones, so that we still find an assimilation effect? These are interesting questions to explore in future research.

Individual perceiver differences

The current work demonstrates the potential impact of perceiver's knowledge about the target on the perceiver's mental representation of the target's face. In this case, knowledge was manipulated through verbal information about the target's behavior, which created an impression about the target's trustworthiness and expected facial features (as shown in Study 3.1). Verbal information about behavior in itself does not create a visual image of a face. What is needed is a perceiver's mental model about how the described behaviors relate to (and are thus used to predict) specific facial features. For example, their mental models may relate the described behaviors to other aspects of the target (e.g. he is probably trustworthy) which may be related to (and thus used to predict) specific facial features. We influenced which mental model became activated through manipulating verbal information. Earlier studies, as described in Chapter 1, usually did this by manipulating visual information (e.g. hairstyle, clothing). We thus rely on participants' associations between the provided information, person impression, and specific facial appearances.

A likely reason that this could lead to perceivable differences in mental representations between experimental conditions is that people generally agree on what facial features make a face appear (un)trustworthy (Oosterhof & Todorov, 2008; Todorov et al., 2013, 2015). In other words, people probably have largely similar mental models on how the described behaviors relate to a target's trustworthiness and how these relate to facial appearance. We took trustworthiness as an example in the present research, but such consensus exists for many more face based impressions than trustworthiness (Todorov, 2017).

However, there are of course also individual differences in perceivers' mental models (Bijlstra et al., 2014; Dotsch et al., 2008; Freeman et al., 2020; Hehman et al., 2019, 2017). Individual perceivers may differ in their (stereotypical) beliefs about targets belonging to certain groups or performing certain types of behaviors. For instance, a target who states to have found God may appear wise and friendly to one perceiver, while unintelligent and annoying to another perceiver. Moreover, individuals may differ in their associations between specific person impressions and specific facial features. For example, a typical criminal may be expected to look sharp vs. thickheaded and of different racial backgrounds depending on the perceiver. Likewise, a mourning target may be expected to be crying in one culture and to be smiling in another.

On top of individual differences in such associations, individuals may also differ in how their associations change dependent on the context (Freeman et al., 2020; Hehman et al., 2019, 2017; Todorov, 2017). For example, for one perceiver, a successful leader may be expected to look dominant during wartime but forgiving during peacetime. For another perceiver, these context-dependent associations could well be different. Relatedly, even if individuals share the same conceptual associations, they may differ in their currently active goals (e.g. to remember a face accurately) and states (e.g. feeling threatened), or in their natural tendencies (e.g. to evaluate others positively or negatively).

Interestingly, such individual differences in conceptual associations, context, goals, states, and natural tendencies may lead to different RIF effects. Future research could pay more attention to the effect of individual perceiver differences, further illuminating to what extent social face perception is biased by perceivers' mental models. Hehman and colleagues (2017) present a promising approach involving cross-classified multilevel models to investigate the unique contribution of perceiver and target characteristics, as well as their interactions, on social face perception.

Does Mental Representation of Face Act as a Mediator?

Amongst others, RIF is societally relevant because specific facial appearances lead to significant social consequences (Antonakis & Eubanks, 2017; Blair et al., 2004; Hassin & Trope, 2000; Olivola & Todorov, 2010a; Porter et al., 2010; Rezlescu et al., 2012; Todorov et al., 2015; van 't Wout & Sanfey, 2008; Wilson &

Rule, 2015; Zebrowitz & McDonald, 1991). The idea is that if verbal information influences the mentally represented seen facial appearance, this in turn should influence those social judgments. In other words, the mental face representation should mediate the effect of verbal information on social consequences. At the same time, it is highly likely that verbal information directly influences social consequences, raising the question how much the indirect effect via mental representations adds, if it adds anything at all. Now that we have established the effect of verbal information on the mental representation of a seen face, future research can investigate the potential (partial) mediating role of such mental representations for the effect of verbal information on social judgments.

Interestingly, research by Kunst and colleagues (2017) suggests a mediating role for mental representations of expected (not seen) US immigrants' facial appearance for the effect of beliefs about the extent to which these immigrants adopted American mainstream culture or maintained their heritage culture on social decisions concerning the probability of racial profiling and qualification for integration support. However, besides investigating mental representations of expected (opposed to seen) faces, the mental face representations and social judgments were not measured in the same sample of participants. The studies used two independent samples of participants: one sample to visualize mental representations of the expected facial appearance (based on beliefs about immigrants' adoption/maintenance of the mainstream/heritage culture) and a different sample to provide social judgments about these visualized mental representations (without any background knowledge about the visualizations). The same holds for research by Ratner and colleagues (2014) on information about minimal group membership, expected facial appearance, and social evaluations and judgments. Future research could provide one and the same sample of participants with verbal information about the target as well as the target's actual face, and measure participants' mental representations of the target's face as well as their social judgment about the target. A mediation analysis on such data may teach us more about the potential mediating role of mental representations of seen faces for the effect of target information on social consequences.

REFLECTIONS ON METHODOLOGICAL AND ANALYTICAL CHOICES

In the present dissertation, we chose to investigate the main research question using online data collection, the data-driven RC methodology, and Bayesian statistical analyses. Here, we critically reflect on and offer some insight into our experiences with these.

Collecting Data Online

Online data collection created some noteworthy benefits and drawbacks. As benefits, it gave us access to large and diverse samples of participants (e.g. concerning residential country, age, educational background), far more diverse than lab studies generally allowed for, increasing the generalizability of our findings. Moreover, we could easily collect data from many participants at the same time, speeding up data collection. As drawbacks, we experienced a server error, which meant we lost data of 19 participants and spent much time e-mailing with participants who had been participating during the server error. To prevent major data losses and large groups of distressed participants under such circumstances, we advise researchers to always collect data in smaller batches, as we did. Additionally, it is impossible to control participants' physical environments, so we had to rely on participants' motivation and willingness to follow our instructions. Unfortunately, a large percentage of participants failed our attention checks. Although we could not use their data, we still had to pay for their participation, costing us resources. We do advise researchers to implement multiple attention checks, so that they can check which participants carefully followed instructions. This can help to estimate and improve data quality (Oppenheimer et al., 2009).

Setting up the experiments online seemed to somewhat compromise their ecological validity at first, but this ecological validity has increased tremendously since the worldwide outbreak of COVID-19. Although people would often meet others in real life, not on a computer screen, in the current lockdowns (Hale et al., 2021), many "live" interactions consist precisely of watching a face on a computer screen. Although we created the experiments online primarily for practical reasons, their online set-up has thus become increasingly relevant to real life since the outbreak of the COVID-19 pandemic.

We made a few other decisions concerning data collection that are worth discussing here. First, we chose to focus on Caucasian participants to study the effect under investigation without potential ingroup/outgroup biases further complicating the research findings. Of course, online data collection gives researchers access to many other ethnic backgrounds as well. Although we have no reasons to suspect that the basic process underlying RIF would work differently for non-Caucasians participants, we certainly encourage researchers to study RIF from the perspective of non-Caucasian participants as well.

Second, we always used a picture of a young adult Caucasian male target person. At the start of our research, we deliberately chose not to spell out the target person's first name (F. Taylor) to maintain the option to select a target face of any race, gender, or age. Sticking with young adult Caucasian males across studies eliminated alternative explanations for the (dis)appearance of RIF in our studies related to variations in the target's gender, race, or age. As such, this choice prevented further complicating the interpretation of our findings. Nevertheless, we consider it a limitation of our research that we only used young adult Caucasian male faces. We therefore encourage researchers to investigate the generalizability of the present RIF effects and their boundary conditions to faces of other races, ages, and genders.

Third, we chose to end the experiment for participants immediately after they failed an attention check, so we could decrease the costs for unusable data (as these participants were paid for their participation up until that point). However, perhaps failing an attention check may not always indicate unmotivated or unserious participation. Although Oppenheimer and colleagues (2009) found no differences between participants who succeeded vs. failed an attention check on age, gender, or need for cognition, there might be other differences. For example, it is conceivable that some of these participants simply have a more impulsive response style in general. Now that online research is becoming increasingly standard, it may be interesting to let such participants finish the whole experiment and investigate how they differ in their responses from participants who succeeded the attention checks. Additionally, participants who failed an attention check could be prompted to try again until they succeed, after which they are allowed to continue with the experiment. This could reduce noise in their responses, improving the quality of their data (Oppenheimer et al., 2009).

Reverse Correlation

As explained in Chapter 1, the RC method is currently considered the best method available to visualize approximations of people's mental representations. Yet, there are some considerations with important theoretical implications that need discussing here.

Do people automatically form visual mental representations?

Study 3.1 showed overwhelming evidence for the Face Prior Hypothesis that verbal information about a person's behavior can generate a visual expectation of that person's face. However, it is possible that the association between participants' conceptual knowledge about the person and specific facial features does not activate automatically, but only when participants are triggered to think about facial appearance, which they obviously are when judging 500 sets of faces in the RC task. Or even more extremely, perhaps participants do not mentally represent the target person's expected face at all, even when triggered to think about facial appearance. Because the RC task forces participants to choose between faces, they may activate prototypical facial appearances based on the verbal information they received only to be able to answer these forced choices, not because that is what they expect the target to look like. If true, the effect might decrease or perhaps even disappear without employment of the RC task. Importantly, this critique holds for all referenced studies investigating expected facial appearances based on social information (Brown-Iannuzzi et al., 2016; Dotsch et al., 2008, 2013; Kunst et al., 2017; Ratner et al., 2014), as they all employed the RC task to measure participants' mental representations.

It is difficult to investigate the automatic nature of a face prior without somehow asking participants about or confronting them with faces. Researchers could ask participants to describe their impression of the target person and investigate how many times participants spontaneously give a description of facial appearance. However, it is possible that participants do not spontaneously mention the facial appearance even when they do have a face prior, either because they think it is not important to mention or because they are not consciously aware of it. Alternatively, researchers could show the actual face without previously asking anything about expected facial appearance and investigate whether participants experience a feeling of 'fit' between the actual face and their expectation of the face (which they may have been unaware of having up until that point).

This could indicate that they did have some sort of face prior before viewing the actual face.

Indeed, we found some indication for such a feeling of (mis)fit in the questionnaire data, especially in Chapter 4. Although we did not enquire this, some participants spontaneously mentioned that the actual face looked differently or similarly to what they expected (e.g. "I remembered that after his description in part 1 that his face wasn't what I expected", "I found it sort of easy to remember his expression because I remember being surprised that I didn't think it fitted with the actions described", "I pictured him as quite a cheerful person but the image I briefly saw showed him as not particularly smiling and with quite a non-expressive facial expression", "I had a preconceived picture in my mind and I was pretty close", "I remember when I saw the original photo of F. Taylor, that he didn't look the way I expected him to – he seemed too grumpy"). Future research could investigate the automatic nature of face priors in a more standardized manner, for instance by having participants indicate on a slider to what extent they experienced surprise or a feeling of fit when viewing the actual face, including an option to indicate they experienced neither. If participants indeed automatically formed a face prior, the RC task can subsequently be used to visualize it.

Given the data from Study 3.1 and the questionnaire data from Chapter 4, it seems likely that verbal information about a person's behavior generates an expectation about his facial appearance. Even if the nature of the RC task enlarges this effect, quotes from participants like the ones mentioned above suggest that the generation of a face prior can indeed occur spontaneously after learning about a person's behavior.

Does the nature of the RC task induce or exaggerate biases?

Once participants have a mental representation of a face, the RC task aims to visualize its most important components. The beauty of the RC task is that it never explicitly mentions the bias researchers are interested in (in this case, facial trustworthiness appearance), but instead measures which facial features participants spontaneously select on. As such, the RC task decreases the chance that it induces biases that would only be present with an explicit prompt. Indeed,

this is one of the main advantages of the RC task over other methods (Brinkman et al., 2017).

However, even for the RC task, it is conceivable that the nature of the task exaggerates or even induces biases in the visualization of the mental representation that are less prominently or not at all present in participants' mental representations. This is possible if there are trials in the RC task on which neither face looks particularly more resembling of the target's actual face to participants. Since participants are forced to choose, they may change their strategy from 'picking the face that looks most like the target's actual face' to 'picking the face that looks most (un)trustworthy', because that fits with the behavioral description of the target person. That way, their response may feel more informed than when they simply choose a face at random, even though they may not believe the target's face to actually look that way. If this happens on enough trials, the resulting averaged CI for this participant will appear slightly (un)trustworthy, even though the participant's mental representation may thus not be biased to the same extent.

Out of the three studies demonstrating a RIF bias, Study 3.4 seems most vulnerable to this problem, because the target's actual face was not included in the base face used in this RC task. Consequently, it is likelier for trials to present two faces that both do not look like the target face to participants. Of course, this was the case also for Study 3.2 and 3.3 in which no RIF effect was found, making it less likely that this suddenly should have become an issue in Study 3.4. Moreover, because the base face was a composite of an average male and average female face, participants had an alternative strategy available: choosing the face that looked more male (opposed to female). As the target person was male in both experimental conditions, this alternative strategy should make it harder, not easier, to find a RIF effect based on the verbal information. Thus, instead of exaggerating a RIF bias, the used base face in Chapter 3 may also have downplayed it.[38]

38 The use of the base face in Chapter 3 allowed us during sequential hypothesis testing to efficiently score participants CIs using the criterion created in Chapter 2. Although this was cost efficient, the fact that the target face was not at all included in the base face thus made it likelier that some trials contained two images that both resembled the target face poorly. For future studies investigating mental representations of a seen face, we advise researchers to use a composite base face that also includes the target's actual face like we did in Chapter 4.

The studies in Chapter 4 used a base face that was a composite of the three different male target faces used in these studies (of which each participant had seen only one face as the target). Consequently, it is likelier that on each trial one face always looked slightly more than the other like the target face that participants had seen. This decreases chances on the problem of a bias induction or exaggeration in these studies. Moreover, it also decreases chances of a bias underestimation as participants can meaningfully focus on which face looks more like the target person opposed to merely more masculine. Interestingly, this could serve as an additional explanation for why it was easier to find a RIF effect in Chapter 4 opposed to 3. Perhaps including the target face in the composite base face increased sensitivity of the RC tasks in Chapter 4 as participants could more meaningfully choose between faces than when neither face resembled the target face, which may have occurred more often in Chapter 3.

To decrease chances on a bias induction, exaggeration, or underestimation even further, a solution for future research might be to investigate the RIF effect using a composite base face that includes the target face as in Chapter 4, but this time employ a four-alternative forced choice (4AFC) RC task instead of a two-images forced choice (2IMF) RC task (Brinkman et al., 2017). Instead of forcing participants to select one out of two images as resembling the target face most, a 4AFC RC task presents participants with only one image per trial and four response options, indicating that the presented face is probably or possibly (not) the target person. To visualize the approximation of the mental representation, only the extreme ('probably') response options are used. That way, images that were uninformative because they did not resemble participants' mental representations at all (and were thus categorized in one of the 'possibly' response options) are discarded and cannot bias the final visualization.

Bayesian Statistical Analyses

We used Bayesian statistical analyses in all studies, which had some noteworthy benefits. First of all, it allowed us to monitor data as they came in (also called sequential hypothesis testing or Sequential Bayes Factors: Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017; Wagenmakers et al., 2018), so that after a preregistered minimum number of participants we could terminate data collection and thus save resources as soon as the data provided convincing evidence for one model over another (which we defined as $1/10 \geqslant BF \geqslant 10$).

Second, by comparing how likely the data were under the predicted vs. the null model, it allowed us to quantify evidence in favor of the null and thus to make claims about the *absence* of an effect, which is impossible under classical null hypothesis significance testing (NHST) with *p*-values (Dienes, 2016; Wagenmakers, Marsman, et al., 2018). Consequently, our studies were not only informative when they provided more evidence in favor of our hypothesis (in comparison to the null), but also when they provided more evidence against it (i.e. in favor of the null), as in Study 3.2.

Another benefit from Bayesian statistics is that researchers can formulate a prior for the alternative model that specifies the distribution of plausible effect sizes the model predicts (Wagenmakers, Marsman, et al., 2018). As Bayes factors quantify the relative probability of the data under the null vs. alternative model, it is important to specify a reasonable alternative model for the comparison to be meaningful. We used the default (and reasonable; van Ravenzwaaij & Wagenmakers, 2019) informed priors from the JASP statistics program (JASP Team, 2020; Wagenmakers, Love, et al., 2018) in all of our studies, never subjective informed priors. Readers may wonder why we did this and whether we should have switched to subjective priors along the way. We outline several reasons for our decision.

Besides the fact that options for subjective priors were limited in JASP at the time we started the RIF studies in early 2017 (Wagenmakers, Love, et al., 2018), the main reason is that we had no clear hypothesis about the size nor direction of the RIF effects. There was no available literature on the RIF effect using the RC task and, as described in Chapter 3, there were theories conceivable that could predict a contrast (opposed to assimilation) effect (Mussweiler, 2003; Rhodes, 2017; Snyder et al., 2015). Researchers should carefully construe their subjective prior choices for the results of their analyses to be meaningful and compelling to other researchers (Wagenmakers, Marsman, et al., 2018). Because we had no strong expectations beforehand, the two-sided default Cauchy prior distribution, which expects smaller effect sizes to be likelier than larger effect sizes, seemed most suitable. This remained the case throughout Chapter 3 as we found convincing evidence for a RIF effect only in the last study of that chapter. In Study 4.1, we used information from Study 3.4 on the direction of the effect (we found an assimilation, not a contrast, effect) to change the default prior

to be one-sided, but left it unchanged otherwise, as we still expected smaller effect sizes to be likelier than larger effect sizes.

Moreover, we did not update priors on effect size (apart from the direction of the effect) based on posteriors from earlier studies because our aim was to investigate the effect of different circumstances on the emergence of RIF (with each study investigating a unique situation), not the overall probability of RIF across circumstances. Each study thus presented a specific new situation for which we investigated afresh whether the RIF effect occurred or not. As research on the circumstances of RIF accumulates and if this accumulation results in strong prior information regarding effect sizes for specific circumstances (van Ravenzwaaij & Wagenmakers, 2019), future studies may build on this accumulation of research findings to inform more subjective priors.

CONCLUSION

In the present dissertation, we aimed to increase understanding of the circumstances under which verbal information about a person's behavior does (not) bias the mental representation of that person's seen face. Throughout the dissertation, we tested two general predictions derived from a Bayesian inspired view on social face perception. In line with the Face Prior Hypothesis, verbal information generated an expectation about the target's facial appearance (i.e. face prior) before seeing the actual face. In line with the Prior-Likelihood Balance Hypothesis, the emergence of a subtle bias in the mental representation of the seen face seemed to depend on the relative strength of the face prior and sensory input of the actual face. The most important factor for the bias to appear seemed to be the accessibility of a person impression of the target (which likely informs the face prior).

The present dissertation demonstrates that the Bayesian inspired view can provide a helpful framework for social face perception researchers to generate hypotheses, specify operationalizations, and organize research findings. Moreover, the Bayesian inspired view helps to reveal the context dependency of the effect of reading into faces. With the present dissertation, we have made an important beginning in uncovering the circumstances under which people are (not) likely to read into faces. We encourage researchers to replicate our

findings and to continue revealing moderating factors of this intriguing bias in social face perception. Once the existence of RIF biases has been established under specific circumstances, researchers can focus their efforts on estimating the strength of such biases and their social consequences.

For now, we can conclude with the nuanced statement that under conditions of strong and accessible expectations, a perceiver's mental image of your face is likely to be slightly biased by his or her expectations about you. When expectations are less accessible, the mental image is more likely to closely resemble your real face. In light of these insights, the common advice to "set expectations low" may not always be the wisest course to steer in our social lives.


A

Appendices

References

REFERENCES

Adams, R. B., Albohn, D. N., & Kveraga, K. (2017). Social Vision: Applying a Social-Functional Approach to Face and Expression Perception. *Current Directions in Psychological Science*, *26*(3), 243–248. https://doi.org/10.1177/0963721417706392

Allport, G. W. (1954). *The nature of prejudice*. Reading: Addison-Wesley Publishing Company. https://doi.org/10.4324/9781912282401

Antonakis, J., & Eubanks, D. L. (2017). Looking Leadership in the Face. *Current Directions in Psychological Science*, *26*(3), 270–275. https://doi.org/10.1177/0963721417705888

Bacchini, F., & Lorusso, L. (2019). Race, again: how face recognition technology reinforces racial discrimination. *Journal of Information, Communication and Ethics in Society*, *17*(3), 321–335. https://doi.org/10.1108/JICES-05-2018-0050

Barden, J., & Tormala, Z. L. (2014). Elaboration and attitude strength: The new meta-cognitive perspective. *Social and Personality Psychology Compass*, *8*(1), 17–29. https://doi.org/10.1111/spc3.12078

Barnett, B. O., Brooks, J. A., & Freeman, J. B. (2020). Stereotypes bias face perception via orbitofrontal–fusiform cortical interaction. *Social Cognitive and Affective Neuroscience*, (September), 1–13. https://doi.org/10.1093/scan/nsaa165

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660. https://doi.org/10.1017/S0140525X99002149

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Bijlstra, G., Holland, R. W., Dotsch, R., Hugenberg, K., & Wigboldus, D. H. J. (2014). Stereotype Associations and Emotion Recognition. *Personality and Social Psychology Bulletin*, *40*(5), 567–577. https://doi.org/10.1177/0146167213520458

Bijlstra, G., Holland, R. W., & Wigboldus, D. H. J. (2010). The social face of emotion recognition: Evaluations versus stereotypes. *Journal of Experimental Social Psychology*, *46*(4), 657–663. https://doi.org/10.1016/j.jesp.2010.03.006

Blair, I. Y., Judd, C. M., & Chapleau, K. M. (2004). 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

Bliss-Moreau, E., Barrett, L. F., & Wright, C. I. (2008). Individual differences in learning the affective value of others under minimal conditions. *Emotion*, *8*(4), 479–493. https://doi.org/10.1037/1528-3542.8.4.479

Bower, G. H. (1991). Mood congruity of social judgments. In J. P. Forgas (Ed.), *International series in experimental social psychology. Emotion and social judgments* (pp. 31–53). New York, US: Pergamon Press.

Bower, G. H., & Karlin, M. B. (1974). Depth of processing pictures of faces and recognition memory. *Journal of Experimental Psychology*, *103*(4), 751–757. https://doi.org/10.1037/h0037190

Brannon, S. M., & Gawronski, B. (2017). A Second Chance for First Impressions? Exploring the Context-(In)Dependent Updating of Implicit Evaluations. *Social Psychological and Personality Science*, *8*(3), 275–283. https://doi.org/10.1177/1948550616673875

Brinkman, L., Todorov, A., & Dotsch, R. (2017). Visualising mental representations: a primer on noise-based reverse correlation in social psychology. *European Review of Social Psychology*, *28*(1), 333–361. https://doi.org/10.1080/10463283.2017.1381469

Brooks, J. A., & Freeman, J. B. (2019). Neuroimaging of person perception: A social-visual interface. *Neuroscience Letters*, *693*(December 2017), 40–43. https://doi.org/10.1016/j.neulet.2017.12.046

Brown-Iannuzzi, J. L., Dotsch, R., Cooley, E., & Payne, B. K. (2016). The Relationship Between Mental Representations of Welfare Recipients and Attitudes Toward Welfare. *Psychological Science*, 1–12. https://doi.org/10.1177/0956797616674999

Bruner, J. S. (1957). On Perception Readiness. *Psychological Review*, *64*(2), 123–152.

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

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–253. https://doi.org/10.1017/S0140525X12000477

Cone, J., Gunaydin, G., & DeLong, J. (2017). *I see a different you: Impression-inconsistent revelations immediately alter mental representations of a person's face*.

Cone, Jeremy, Brown-Iannuzzi, J. L., Lei, R., & Dotsch, R. (2020). Type I Error Is Inflated in the Two-Phase Reverse Correlation Procedure. *Social Psychological and Personality Science*, 1–9. https://doi.org/10.1177/1948550620938616

Cone, Jeremy, & Ferguson, M. J. (2015). He Did What ? The Role of Diagnosticity in Revising Implicit Evaluations. *Journal of Personality and Social Psychology*, *108*(1), 37–57. https://doi.org/10.1037/pspa0000014

Cone, Jeremy, Mann, T. C., & Ferguson, M. J. (2017). *Changing Our Implicit Minds: How, When, and Why Implicit Evaluations Can Be Rapidly Revised*. *Advances in Experimental Social Psychology* (1st ed.). Elsevier Inc. https://doi.org/10.1016/bs.aesp.2017.03.001

De Houwer, J. (2018). Propositional Models of Evaluative Conditioning. *Social Psychological Bulletin*, *13*(3), 1–21. https://doi.org/10.5964/spb.v13i3.28046

DeBruine, L. (2017). Webmorph (Version v0.0.0.9000). *Zenodo*. https://doi.org/10.5281/zenodo.1073697

DeBruine, L. M. (2002). Facial resemblance enhances trust. *Proceedings of the Royal Society B: Biological Sciences*, *269*(1498), 1307–1312. https://doi.org/10.1098/rspb.2002.2034

Deffenbacher, K. A., Bornstein, B. H., McGorty, E. K., & Penrod, S. D. (2008). Forgetting the once-seen face: Estimating the strength of an eyewitness's memory representation. *Journal of Experimental Psychology: Applied*, *14*(2), 139–150. https://doi.org/10.1037/1076-898X.14.2.139

Devine, P. G. (1989). Stereotypes and Prejudice: Their Automatic and Controlled Components. *Journal of Personality and Social Psychology*, *56*(1), 5–18.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003

Dotsch, R. (2016). rcicr: Reverse correlation image classification toolbox. R package version 0.3.4.1.

Dotsch, R. (2017). rcicr: Reverse correlation image classification toolbox. R package version 0.4.0.

Dotsch, R., Hassin, R. R., & Todorov, A. (2016). Statistical learning shapes face evaluation. *Nature Human Behaviour*, *1*(November), 1–6. https://doi.org/10.1038/s41562-016-0001

Dotsch, R., & Todorov, A. (2012). Reverse Correlating Social Face Perception. *Social Psychological and Personality Science*, *3*(5), 562–571. https://doi.org/10.1177/1948550611430272

Dotsch, R., Wigboldus, D. H. J., Langner, O., & Van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, *19*(10), 978–980. https://doi.org/10.1111/j.1467-9280.2008.02186.x

Dotsch, R., Wigboldus, D. H. J., & Van Knippenberg, A. (2013). Behavioral information biases the expected facial appearance of members of novel groups. *European Journal of Social Psychology*, *43*(1), 116–125. https://doi.org/10.1002/ejsp.1928

Dunbar, R. I. M., Marriott, A., & Duncan, N. D. C. (1997). Human conversational behavior. *Human Nature*, *8*(3), 231–246. https://doi.org/10.1007/BF02912493

Edwards, M. J., Adams, R. A., Brown, H., Pareés, I., & Friston, K. J. (2012). A Bayesian account of "hysteria." *Brain*, *135*(11), 3495–3512. https://doi.org/10.1093/brain/aws129

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

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. https://doi.org/10.1038/415429a

Falvello, V., Vinson, M., Ferrari, C., & Todorov, A. (2015). The Robustness of Learning about the Trustworthiness of Other People. *Social Cognition*, *33*(5), 368–386. https://doi.org/10.1521/soco.2015.33.5.368

Fazio, R. H. (2007). Attitudes as Object-Evaluation Associations of Varying Strength. *Social Cognition*, *25*(5), 603–637. https://doi.org/10.1521/soco.2007.25.5.603.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the Automatic Activation of Attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238. https://doi.org/10.1037/0022-3514.50.2.229

Fetsch, C. R., Pouget, A., DeAngelis, G. C., & Angelaki, D. E. (2012). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience*, *15*(1), 146–154. https://doi.org/10.1038/nn.2983

Fiedler, K., Kaczor, K., Haarmann, S., Stegmüller, M., & Maloney, J. (2009). Impression-formation advantage in memory for faces: When eyewitnesses are interested in targets' likeability, rather than their identity. *European Journal of Social Psychology*, *39*, 793–807. https://doi.org/10.1002/ejsp.581

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, *47*(6), 1231–1236. https://doi.org/10.1016/j.jesp.2011.05.007

Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception*, *37*(4), 571–583. https://doi.org/10.1068/p5705

Flowe, H. D., & Humphries, J. E. (2011). An examination of criminal face bias in a random sample of police lineups. *Applied Cognitive Psychology*, *25*(2), 265–273. https://doi.org/10.1002/acp.1673

Förderer, S., & Unkelbach, C. (2013). On the stability of evaluative conditioning effects : The role of identity memory, valence memory, and evaluative consolidation. *Social Psychology*, *44*(6), 380–389. https://doi.org/10.1027/1864-9335/a000150

Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M., & Ambady, N. (2011). Looking the part: Social status cues shape race perception. *PLoS ONE*, *6*(9). https://doi.org/10.1371/journal.pone.0025107

Freeman, J. B., Stolier, R. M., & Brooks, J. A. (2020). *Dynamic interactive theory as a domain-general account of social perception. Advances in Experimental Social Psychology* (1st ed., Vol. 61). Elsevier Inc. https://doi.org/10.1016/bs.aesp.2019.09.005

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Fuhrman, R. W., Bodenhausen, G. V., & Lichtenstein, M. (1989). On the trait implications of social behaviors: Kindness, intelligence, goodness, and normality ratings for 400 behavior statements. *Behavior Research Methods, Instruments, & Computers*, *21*(6), 587–597. https://doi.org/10.3758/BF03210581

Galperin, A., & Haselton, M. G. (2013). Error Management and the evolution of cognitive bias. In J. P. Forgas, K. Fiedler, & C. Sedikides (Eds.), *Social Thinking and Interpersonal Behavior* (pp. 45–63). Psychology Press. https://doi.org/10.4324/9780203139677

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal Predictions in Everyday Cognition. *Psychological Science*, *17*(9), 767–773.

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., … Tatlow, H. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*, *5*, 529–538. https://doi.org/10.1038/s41562-021-01079-8

Hamilton, D. L., Katz, L. B., & Leirer, V. O. (1980). Cognitive representation of personality impressions: Organizational processes in first impression formation. *Journal of Personality and Social Psychology*, *39*(6), 1050–1063. https://doi.org/10.1037/h0077711

Hassin, R. R., Aviezer, H., & Bentin, S. (2013). Inherently ambiguous: Facial expressions of emotions, in context. *Emotion Review*, *5*(1), 60–65. https://doi.org/10.1177/1754073912451331

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., 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

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687–701. https://doi.org/10.1016/j.cognition.2008.05.010

Houston, D. A., & Fazio, R. H. (1989). Biased Processing as a Function of Attitude Accessibility: Making Objective Judgments Subjectively. *Social Cognition*, *7*(1), 51–66. https://doi.org/10.1521/soco.1989.7.1.51

Huang, J. Y., & Bargh, J. A. (2014). The Selfish Goal: Autonomously operating motivational structures as the proximate cause of human judgment and behavior. *Behavioral and Brain Sciences*, *38*(01), 121–135. https://doi.org/10.1017/S0140525X13000290

Hugenberg, K. (2005). Social categorization and the perception of facial affect: Target race moderates the response latency advantage for happy faces. *Emotion*, *5*(3), 267–276. https://doi.org/10.1037/1528-3542.5.3.267

Hutmacher, F. (2019). Why Is There So Much More Research on Vision Than on Any Other Sensory Modality? *Frontiers in Psychology*, *10*, 1–12. https://doi.org/10.3389/fpsyg.2019.02246

Imhoff, R., & Dotsch, R. (2013). Do We Look Like Me or Like Us? Visual Projection as Self- or Ingroup-Projection. *Social Cognition*, *31*(6), 806–816. https://doi.org/10.1521/soco.2013.31.6.806

Jack, R. E., & Schyns, P. G. (2017). Toward a Social Psychophysics of Face Communication. *Annual Review of Psychology*, *68*(1), 269–297. https://doi.org/10.1146/annurev-psych-010416-044242

Jackson, R. E., & Cormack, L. K. (2007). Evolved navigation theory and the descent illusion. *Perception and Psychophysics*, *69*(3), 353–362. https://doi.org/10.3758/BF03193756

JASP Team. (2020). JASP (Version 0.14.1) [Computer software].

Jeffreys, H. (1961). *Theory of Probability. Theory of Probability* (Vol. 2). Retrieved from http://ocw.mit.edu/OcwWeb/Mathematics/18-175Spring-2007/LectureNotes/Index.htm

Johnson, K. L., Freeman, J. B., & Pauker, K. (2012). Race is gendered: How covarying phenotypes and stereotypes bias sex categorization. *Journal of Personality and Social Psychology*, *102*(1), 116–131. https://doi.org/10.1037/a0025335

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, *13*(2), 150–158. https://doi.org/10.1016/S0959-4388(03)00042-4

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing*, *8*(3), 159–166. https://doi.org/10.1007/s10339-007-0170-2

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

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kunst, J. R., Dovidio, J. F., & Dotsch, R. (2017). White Look-Alikes: Mainstream Culture Adoption Makes Immigrants "Look" Phenotypically White. *Personality and Social Psychology Bulletin*, 014616721773927. https://doi.org/10.1177/0146167217739279

Lammers, J., Gast, A., Unkelbach, C., & Galinsky, A. D. (2017). Moral Character Impression Formation Depends on the Valence Homogeneity of the Context. *Social Psychological and Personality Science*, 1–10. https://doi.org/10.1177/1948550617714585

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

Levin, D. T., & Banaji, M. R. (2006). Distortions in the perceived lightness of faces: The role of race categories. *Journal of Experimental Psychology: General*, *135*(4), 501–512. https://doi.org/10.1037/0096-3445.135.4.501

Lippmann, W. (1922). *Public opinion*. New York, US: Harcourt, Brace, and Company Inc.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section. *Karolinska Institut, ISBN 91-630-7164-9*. https://doi.org/10.1017/CBO9781107415324.004

Lundqvist, D., & Litton, J. E. (1998). The Averaged Karolinska Directed Emotional Faces - AKDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.

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

Mamassian, P., Landy, M., & Maloney, L. T. (2002). Bayesian modelling of visual perception. In R. P. N. Rao, B. A. Olshausen, & M. S. Lewicki (Eds.), *Probabilistic models of the brain: Perception and neural function* (pp. 13–36). Cambridge, Massachusetts: MIT Press.

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science*, *28*(2), 209–226. https://doi.org/10.1016/j.cogsci.2003.11.004

Marzi, T., Righi, S., Ottonello, S., Cincotta, M., & Viggiano, M. P. (2014). Trust at first sight: Evidence from ERPs. *Social Cognitive and Affective Neuroscience*, *9*(1), 63–72. https://doi.org/10.1093/scan/nss102

Mattarozzi, K., Todorov, A., & Codispoti, M. (2015). Memory for faces: the effect of facial appearance and the context in which the face is encountered. *Psychological Research*, *79*(2), 308–317. https://doi.org/10.1007/s00426-014-0554-8

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*(6), 255–260. Retrieved from http://tics.trends.com1364-6613/02/$-seefrontmatter

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say "hello"? Personality impressions from brief novel voices. *PLoS ONE*, *9*(3), 1–9. https://doi.org/10.1371/journal.pone.0090779

Medvec, V. H., Madey, S. F., & Gilovich, T. (1995). When Less Is More: Counterfactual Thinking and Satisfaction among Olympic Medalists. *Journal of Personality and Social Psychology*, *69*(4), 603–610. https://doi.org/10.1017/cbo9780511808098.037

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2002). Optimal methods for calculating classification images: Weighted sums. *Journal of Vision*, *2*(1), 79–104. https://doi.org/10.1167/2.1.6

Mussweiler, T. (2003). Comparison processes in social judgment: mechanisms and consequences. *Psychological Review*, *110*(3), 472–489. https://doi.org/10.1037/0033-295X.110.3.472

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 of the United States of America*, *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

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Ouyang, S., Hospedales, T. M., Song, Y. Z., & Li, X. (2016). ForgetMeNot: Memory-aware forensic facial sketch matching. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 5571–5579. https://doi.org/10.1109/CVPR.2016.601

Pizarro, D. A., Laney, C., Morris, E. K., & Loftus, E. F. (2006). Ripple effects in memory: judgments of moral blame can distort memory for events. *Memory & Cognition*, *34*(3), 550–555. https://doi.org/10.3758/BF03193578

Plant, E. A., & Peruche, B. M. (2005). The consequences of race for police officers' responses to criminal suspects. *Psychological Science*, *16*(3), 180–183. https://doi.org/10.1111/j.0956-7976.2005.00800.x

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

Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning – induced hallucinations result from overweighting of perceptual priors. *Science*, *357*(6351), 596–600. https://doi.org/10.1126/science.aan3458

Quinn, K. A., Macrae, C. N., & Bodenhausen, G. V. (2007). Stereotyping and impression formation: How categorical thinking shapes person perception. In M. A. Hogg & J. Cooper (Eds.), *The SAGE Handbook of Social Psychology: Concise Student Edition* (pp. 68–92). London: SAGE Publications Ltd. https://doi.org/10.4135/9781848608221.n4

R Development Core Team. (2016). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing Vienna Austria*, *0*, {ISBN} 3-900051-07-0. https://doi.org/10.1038/sj.hdy.6800737

Ratner, K. G., Dotsch, R., Wigboldus, D. H. J., van Knippenberg, A., & Amodio, D. M. (2014). Visualizing minimal ingroup and outgroup faces: implications for impressions, attitudes, and behavior. *Journal of Personality and Social Psychology*, *106*(6), 897–911. https://doi.org/10.1037/a0036498

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), 1–6. https://doi.org/10.1371/journal.pone.0034293

Rhodes, G. (2017). Adaptive Coding and Face Recognition. *Current Directions in Psychological Science*, *26*(3), 218–224. https://doi.org/10.1177/0963721417692786

Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision Research*, *46*(18), 2977–2987. https://doi.org/10.1016/j.visres.2006.03.002

Rule, N. O., & Sutherland, S. L. (2017). Social Categorization From Faces: Evidence From Obvious and Ambiguous Groups. *Current Directions in Psychological Science*, *26*(3), 231–236. https://doi.org/10.1177/0963721417697970

San Roque, L., Kendrick, K. H., Norcliffe, E., Brown, P., Defina, R., Dingemanse, M., … Majid, A. (2015). Vision verbs dominate in conversation across cultures, but the ranking of non-visual verbs varies. *Cognitive Linguistics*, *26*(1), 31–60. https://doi.org/10.1515/cog-2014-0089

Sanbonmatsu, D. M., & Fazio, R. H. (1990). The Role of Attitudes in Memory-Based Decision Making. *Journal of Personality and Social Psychology*, *59*(4), 614–622. https://doi.org/10.1037/0022-3514.59.4.614

Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. https://doi.org/10.1037/met0000061

Seifert, L. S. (1997). Activating representations in permanent memory: Different benefits for pictures and words. *Journal of Experimental Psychology: Learning Memory and Cognition*, *23*(5), 1106–1121. https://doi.org/10.1037/0278-7393.23.5.1106

Shepard, R. N., & Metzler, J. (1973). Mental rotation of three-dimensionsal objects. *Science, 171*(3972), 701-703.

Shepherd, J. W., Gibling, F., & Ellis, H. D. (1991). The Effects of Distinctiveness, Presentation Time and Delay on Face Recognition. *European Journal of Cognitive Psychology*, *3*(1), 137–145. https://doi.org/10.1080/09541449108406223

Sherman, J. W., Lee, A. Y., Bessenoff, G. R., & Frost, L. A. (1998). Stereotype Efficiency Reconsidered: Encoding Flexibility under Cognitive Load. *Journal of Personality and Social Psychology*, *75*(3), 589–606. https://doi.org/10.1037/0022-3514.75.3.589

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18839484

Snyder, J. S., Schwiedrzik, C. M., Vitela, A. D., & Melloni, L. (2015). How previous experience shapes perception in different sensory modalities. *Frontiers in Human Neuroscience*, *9*(October), 1–8. https://doi.org/10.3389/fnhum.2015.00594

Sofer, C., Dotsch, R., Wigboldus, D. H. J., & Todorov, A. (2015). What Is Typical Is Good: The Influence of Face Typicality on Perceived Trustworthiness. *Psychological Science*, *26*(1), 39–47. https://doi.org/10.1177/0956797614554955

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*(9), 403–409. https://doi.org/10.1016/j.tics.2009.06.003

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

Todorov, A. (2017). *Face Value: The Irresistible Influence of First Impressions*. Princeton and Oxford: Princeton University Press.

Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion*, *13*(4), 724–738. https://doi.org/10.1037/a0032335

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*, 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., & Oosterhof, N. (2011). Modeling Social Perception of Faces. *IEEE Signal Processing Magazine*, *28*(2), 117–122. https://doi.org/10.1109/MSP.2010.940006

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., 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

Trevethan, R. (2017). Intraclass correlation coefficients: clearing the air, extending some cautions, and making some requests. *Health Services and Outcomes Research Methodology*, *17*(2), 127–143. https://doi.org/10.1007/s10742-016-0156-6

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

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217–239. https://doi.org/10.1037/met0000100

Van den Stock, J., Righart, R., & de Gelder, B. (2007). Body Expressions Influence Recognition of Emotions in the Face and Voice. *Emotion*, *7*(3), 487–494. https://doi.org/10.1037/1528-3542.7.3.487

van Ravenzwaaij, D., & Wagenmakers, E.-J. (2019). *Advantages Masquerading as 'Issues' in Bayesian Hypothesis Testing: A Commentary on Tendeiro and Kiers (2019)*. https://doi.org/10.31234/osf.io/nf7rp

Wagenaar, W. A. (1989). *Het herkennen van Iwan: de identificatie van de dader door ooggetuigen van een misdrijf*. Lisse: Swets & Zeitlinger.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin and Review*, *25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., … Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wheeler, M. E., Petersen, S. E., & Buckner, R. L. (2000). Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(20), 11125–11129.

Wigboldus, D. H. J., & Dotsch, R. (2016). Encourage Playing with Data and Discourage Questionable Reporting Practices. *Psychometrika*, *81*(1), 27–32. https://doi.org/10.1007/s11336-015-9445-1

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., Hugenberg, K., & Rule, N. O. (2017). Racial bias in judgments of physical size and formidability: From size to threat. *Journal of Personality and Social Psychology*, *113*(1), 59–80. https://doi.org/10.1037/pspi0000092

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

Wincenciak, J., Dzhelyova, M., Perrett, D. I., & Barraclough, N. E. (2013). Adaptation to facial trustworthiness is different in female and male observers. *Vision Research*, *87*, 30–34. https://doi.org/10.1016/j.visres.2013.05.007

Zaki, J. (2013). Cue Integration: A Common Framework for Social Cognition and Physical Perception. *Perspectives on Psychological Science*, *8*(3), 296–312. https://doi.org/10.1177/1745691613475454

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., & 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

Supplemental materials Chapter 3

'Under Which Circumstances Does Non-Visual Behavioral Information (Not) Influence Visual Mental Representations of Seen Faces?'

Table S3.1. Number of participants per country included in the analyses of each study. Country information is based on participants' self-reported current country of residence.

	Number of participants (percentage of study sample)			
Country	**Study 3.1**	**Study 3.2**	**Study 3.3**	**Study 3.4**
Australia	-	-	1 (0.50%)	1 (0.50%)
Austria	-	-	-	1 (0.50%)
Belgium	-	2 (1.18%)	-	1 (0.50%)
Bosnia and Herzegovina	-	11 (6.47%)	-	-
Brazil	-	-	1 (0.50%)	-
Bulgaria	1 (2.50%)	8 (4.71%)	1 (0.50%)	-
Canada	2 (5.00%)	2 (1.18%)	5 (2.50%)	1 (0.50%)
Croatia	-	3 (1.77%)	-	-
Czech Republic	-	-	2 (1.00%)	3 (1.50%)
Denmark	-	-	1 (0.50%)	1 (0.50%)
Finland	-	-	4 (2.00%)	1 (0.50%)
France	-	2 (1.18%)	1 (0.50%)	3 (1.50%)
Germany	2 (5.00%)	3 (1.77%)	3 (1.50%)	1 (0.50%)
Greece	1 (2.50%)	1 (0.59%)	4 (2.00%)	2 (1.00%)
Hungary	-	1 (0.59%)	4 (2.00%)	1 (0.50%)
India	-	1 (0.59%)	-	-
Ireland	1 (2.50%)	2 (1.18%)	2 (1.00%)	1 (0.50%)
Italy	1 (2.50%)	4 (2.35%)	6 (3.00%)	4 (2.00%)
Japan	-	-	-	1 (0.50%)
Latvia	-	1 (0.59%)	-	-
Macedonia	-	3 (1.77%)	-	-
Mexico	-	1 (0.59%)	-	-
Netherlands	1 (2.50%)	4 (2.35%)	1 (0.50%)	1 (0.50%)
Poland	-	2 (1.18%)	7 (3.50%)	7 (3.50%)
Portugal	-	9 (5.29%)	10 (5.00%)	14 (7.00%)
Romania	-	1 (0.59%)	1 (0.50%)	-
Serbia	-	26 (15.29%)	1 (0.50%)	-
Slovenia	-	1 (0.59%)	1 (0.50%)	1 (0.50%)
South Africa	-	1 (0.59%)	-	-
South Korea	-	-	-	1 (0.50%)

Table S3.1. Number of participants per country included in the analyses of each study. Country information is based on participants' self-reported current country of residence. (continued)

	Number of participants (percentage of study sample)			
Country	**Study 3.1**	**Study 3.2**	**Study 3.3**	**Study 3.4**
Spain	-	3 (1.77%)	1 (0.50%)	4 (2.00%)
Sweden	1 (2.50%)	3 (1.77%)	1 (0.50%)	1 (0.50%)
Switzerland	-	-	-	1 (0.50%)
United Kingdom	22 (55.00%)	59 (34.71%)	100 (50.00%)	129 (64.50%)
United States	8 (20.00%)	16 (9.41%)	42 (21.00%)	19 (9.50%)
Total	**40 (100%)**	**170 (100%)**	**200 (100%)**	**200 (100%)**

Table S3.2. Number of participants per highest completed education level included in the analyses of each study.

	Number of participants (percentage of study sample)			
Highest completed education	**Study 3.1**	**Study 3.2**	**Study 3.3**	**Study 3.4**
Primary education	-	4 (2.35%)	5 (2.50%)	1 (0.50%)
Lower secondary education	-	5 (2.94%)	3 (1.50%)	8 (4.00%)
Upper secondary education	6 (15.00%)	31 (18.24%)	45 (22.50%)	39 (19.50%)
Post-secondary non-tertiary education	3 (7.50%)	12 (7.06%)	11 (5.50%)	27 (13.50%)
Short-cycle tertiary education	6 (15.00%)	8 (4.71%)	13 (6.50%)	10 (5.00%)
Bachelor's or equivalent level	16 (40.00%)	70 (41.18%)	81 (40.50%)	77 (38.50%)
Master's or equivalent level	8 (20.00%)	35 (20.59%)	36 (18.00%)	34 (17.00%)
Doctoral or equivalent level	1 (2.50%)	5 (2.94%)	6 (3.00%)	4 (2.00%)
Total	**40 (100%)**	**170 (100%)**	**200 (100%)**	**200 (100%)**

APPENDIX S-3A

At the end of each study, participants answered the following questions about their experience during the experiment.

Study 3.1

1. How hard they found it to remember the behaviors of F. Taylor on a 5-point scale from 1 (*very easy*) to 5 (*very hard*). Participants were free to leave additional comments.
2. To describe in their own words what they expected F. Taylor's face to look like.
3. How hard they found it to each time select the face they would say was most likely F. Taylor on a 5-point scale from 1 (*very easy*) to 5 (*very hard*). Participants were free to leave additional comments.
4. What characteristics of the faces presented participants based their decisions on.
5. To what extent participants believed one can know an individual's true personality from looking at the face (as formulated by Hassin and Trope, 2000).

Study 3.2

1. How hard they found it to remember the behaviors of F. Taylor on a 5-point scale from 1 (*very easy*) to 5 (*very hard*). Participants were free to leave additional comments.
2. How hard they found it to each time select the face that looked most like F. Taylor on a 5-point scale from 1 (*very easy*) to 5 (*very hard*). Participants were free to leave additional comments.
3. What characteristics of the faces presented participants based their decisions on.

Study 3.3

1. How hard they found it to remember the behaviors of F. Taylor on a 5-point scale from 1 (*very easy*) to 5 (*very hard*). Participants were free to leave additional comments.
2. Whether they had seen F. Taylor's face at all (yes / no). Participants were free to leave additional comments.

3. How hard they found it to each time select the face that looked most like F. Taylor to them on a 5-point scale from 1 (*very easy*) to 5 (*very hard*). Participants were free to leave additional comments.
4. What characteristics of the faces presented participants based their decisions on.

Study 3.4

1. How hard they found it to remember the behaviors of F. Taylor on a 5-point scale from 1 (*very easy*) to 5 (*very hard*). Participants were free to leave additional comments.
2. Whether they had seen F. Taylor's face at all (yes / no). Participants were free to leave additional comments.
3. How hard they found it to each time select the face that looked most like F. Taylor to them on a 5-point scale from 1 (*very easy*) to 5 (*very hard*). Participants were free to leave additional comments.
4. What characteristics of the faces presented participants based their decisions on.
5. To what extent participants believed one can know an individual's true personality from looking at the face (as formulated by Hassin and Trope, 2000).

APPENDIX S-3B

In Study 3.3, 10 participants indicated they had not seen F. Taylor's face during the face presentation task. We generated the group CIs with and without the individual CIs of these 10 participants to inspect on potential differences that may have driven the effect we found on the trustworthiness ratings of the group CIs (Figure S-3B1). On visual inspection, including or excluding these individual CIs does not appear to make a difference. This suggests that the effect of behavioral condition (untrustworthy / trustworthy) on the trustworthiness ratings of the group CIs of Study 3.3 is not driven by the individual CIs of the 10 participants who indicated that they had not seen F. Taylor's face.

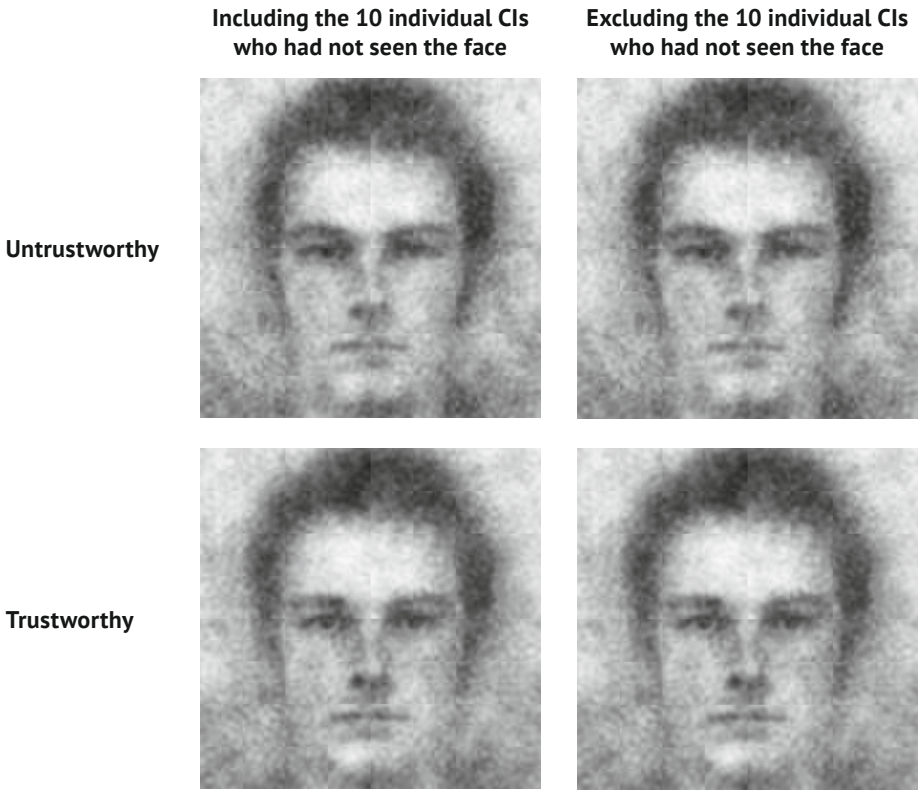


Figure S-3B1. Study 3.3's group CIs of F. Taylor's facial appearance in the untrustworthy (top row) and trustworthy (bottom row) conditions including (left column) and excluding (right column) the individual CIs of the 10 participants who indicated they had not seen F. Taylor's face during the face presentation task.

In Study 3.4, 8 participants indicated they had not seen F. Taylor's face during the face presentation task. Here too, we generated the group CIs with and without the individual CIs of these 8 participants to inspect on potential differences that may have driven the effect we found on the trustworthiness ratings of the group CIs (Figure S-3B2). Again, including or excluding these individual CIs did not lead to clear visual changes.

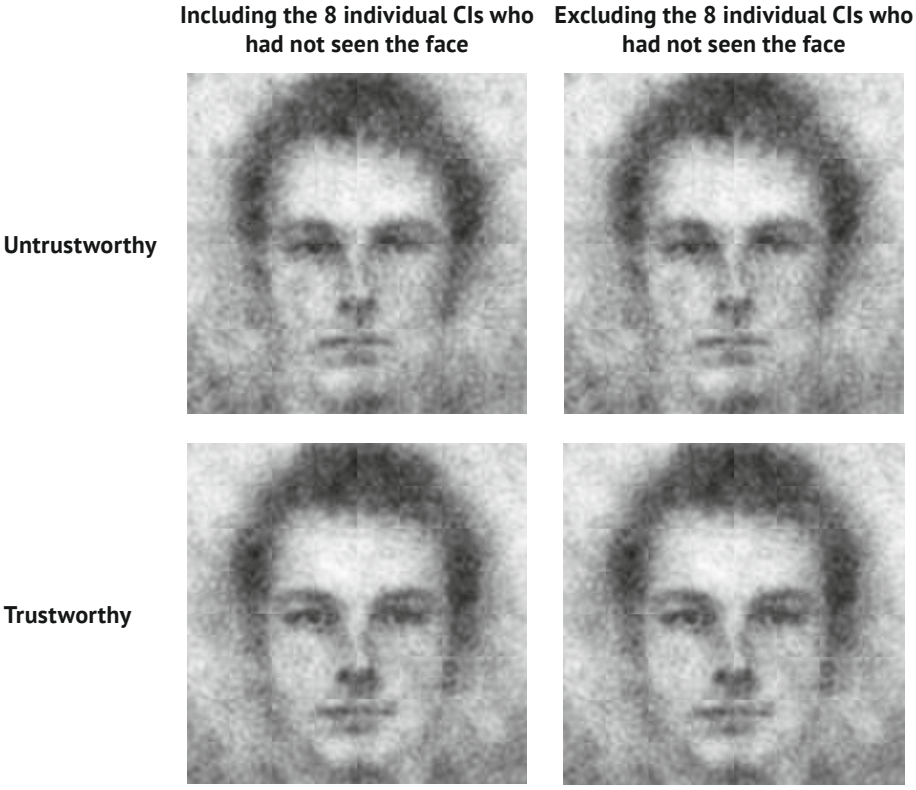
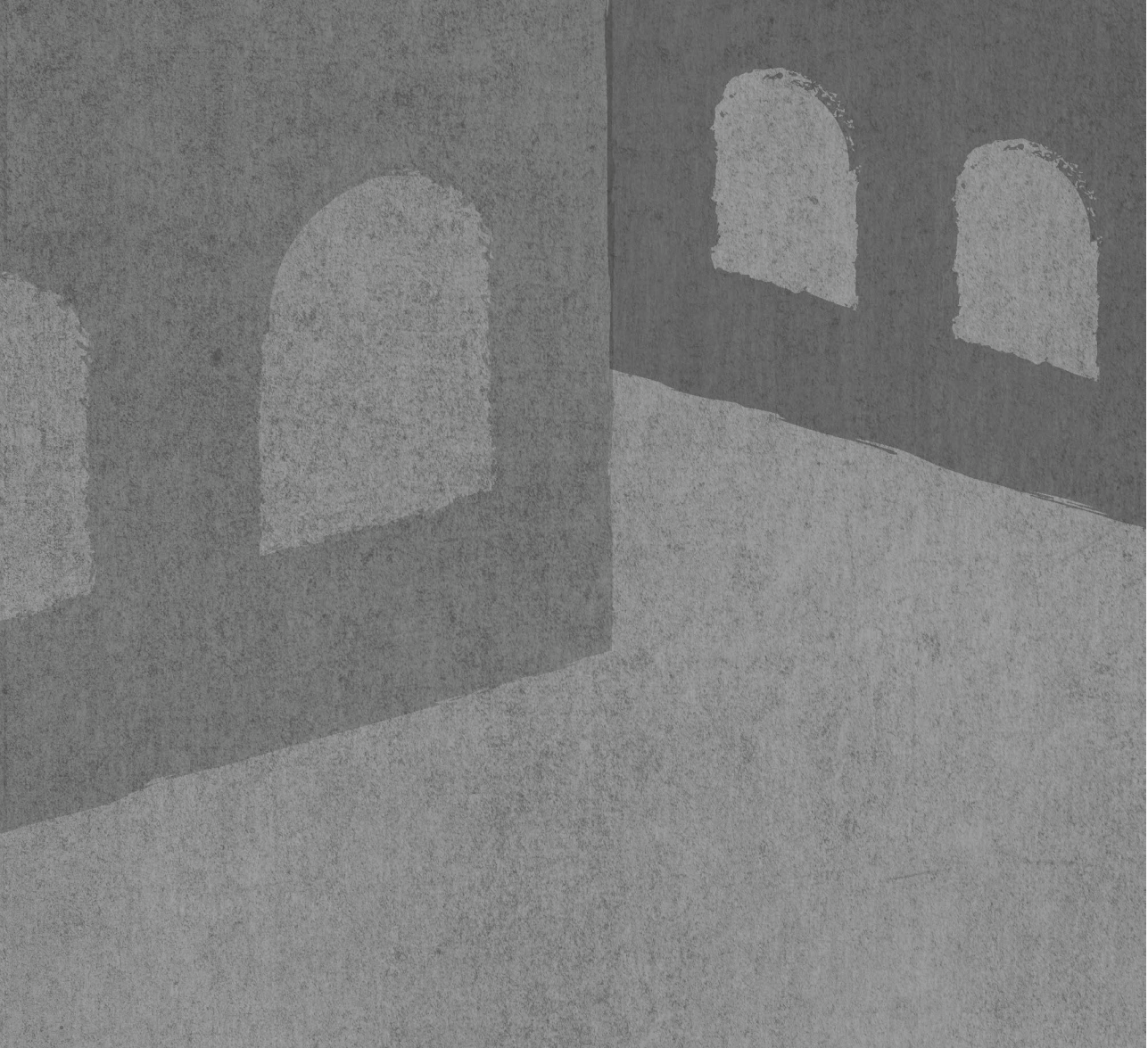


Figure S-3B2. Study 3.4's group CIs of F. Taylor's facial appearance in the untrustworthy (top row) and trustworthy (bottom row) conditions including (left column) and excluding (right column) the individual CIs of the 8 participants who indicated they had not seen F. Taylor's face during the face presentation task.



Supplemental materials Chapter 4

'Temporal Stability of Biases in Mental Representations of Faces'

Table S4.1. Number of participants per self-reported current country of residence included in the analyses of each study.

	Number of participants (percentage of study sample)	
Country	**Study 4.1**	**Study 4.2**
Australia	2 (0.67%)	4 (1.33%)
Austria	1 (0.33%)	1 (0.33%)
Belgium	1 (0.33%)	1 (0.33%)
Canada	7 (2.33%)	3 (1.00%)
Czech Republic	2 (0.67%)	1 (0.33%)
Denmark	-	1 (0.33%)
Estonia	2 (0.67%)	1 (0.33%)
Finland	1 (0.33%)	1 (0.33%)
France	3 (1.00%)	-
Germany	3 (1.00%)	5 (1.67%)
Greece	2 (0.67%)	5 (1.67%)
Hungary	4 (1.33%)	2 (0.67%)
Ireland	2 (0.67%)	1 (0.33%)
Israel	1 (0.33%)	-
Italy	13 (4.33%)	9 (3.00%)
Malta	1 (0.33%)	-
Netherlands	4 (1.33%)	3 (1.00%)
Poland	4 (1.33%)	5 (1.67%)
Portugal	15 (5.00%)	9 (3.00%)
Slovenia	-	1 (0.33%)
Spain	9 (3.00%)	3 (1.00%)
Turkey	2 (0.67%)	-
United Kingdom	202 (67.33%)	203 (67.67%)
United States	19 (6.33%)	41 (13.67%)
Total	**300 (100%)**	**300 (100%)**

Table S4.2. Number of participants per self-reported highest completed education level included in the analyses of each study.

	Number of participants (percentage of study sample)	
Highest completed education	**Study 4.1**	**Study 4.2**
Primary education	1 (0.33%)	2 (0.67%)
Lower secondary education	7 (2.33%)	8 (2.67%)
Upper secondary education	58 (19.33%)	51 (17.00%)
Post-secondary non-tertiary education	46 (15.33%)	29 (9.67%)
Short-cycle tertiary education	8 (2.67%)	25 (8.33%)
Bachelor's or equivalent level	120 (40.00%)	136 (45.33%)
Master's or equivalent level	48 (16.00%)	42 (14.00%)
Doctoral or equivalent level	12 (4.00%)	7 (2.33%)
Total	**300 (100%)**	**300 (100%)**

APPENDIX S-4A

Participants answered the following questions about their experience during the experiment at the end of each study. The fifth question was asked in Study 4.1 only.

1. How hard they found it to remember their impression of F. Taylor on a 5-point scale from 1 (*very easy*) to 5 (*very hard*). Participants were free to leave additional comments.
2. How hard they found it to each time select the face that looked most like F. Taylor to them on a 5-point scale from 1 (*very easy*) to 5 (*very hard*). Participants were free to leave additional comments.
3. What characteristics of the faces presented participants based their decisions on.
4. To what extent participants believed one can know an individual's true personality from looking at the face (as formulated by Hassin and Trope, 2000).
5. How many nights of sleep participants had had since they finished part 1 of this study.

English summary

Faces play a highly significant role in people's social lives. Not only do faces receive most attention in everyday social interactions, they also have a profound influence on how the perceiver thinks, feels, and behaves toward the face bearer, impacting social decisions that carry consequences which go far beyond that specific social interaction. For instance, scientific studies in social face perception have shown that facial appearance can influence decisions as consequential as whether or not to vote for someone in politics, hire someone for a job, trust someone financially, or even sentence a defendant to death.

Interestingly though, people's cognitive systems do not operate like passive recorders, but rather actively construct their own mental representations of reality based on the sensory input they receive form the world "out there" and their available knowledge about the world. This raises the question how accurately people mentally represent others' faces, especially when they already have some knowledge about the other person. Besides influencing social evaluations, then, could mental representations of faces themselves also be influenced by social information? If so, people do not only read information *from* faces, but also *into* faces.

Indeed, scientific studies have shown that social visual cues, such as hairstyle, clothing, gender and race appearance, can influence social categorization of a face, for instance on race, gender, and emotional expression. But what about social cues that are not present in the visual field at the same time as the face? Could social knowledge based on verbal information influence the mental representation of a seen face as well? If so, this would substantiate the theoretical idea that mental representations of faces can be influenced both by sensory input from the face that is "out there" and by other social beliefs the perceiver might have about the face bearer. Moreover, it would imply that the potential power to influence others' mental representations of a target's face lies not just with the target person and the select few who may alter his/her visual appearance (say, the hairdresser), but in fact with everyone who can share credible appearing information about the target (e.g. through gossip).

Understanding whether and under which circumstances verbal information about a person may influence others' mental representations of the person's seen face is therefore both theoretically and societally relevant. It can increase

our theoretical understanding of the extent and circumstances in which our cognitive systems use available social information that is not visually present to inform a mental representation of a person's face. Moreover, such understanding can raise societal awareness of the circumstances under which people's mental representations of other's faces are likely to be colored.

The aim of the current dissertation was to increase understanding of the circumstances under which verbal information about a person's behavior is (not) likely to bias a perceiver's visual mental representation of that person's once seen face, also called 'reading into faces' (RIF). Besides investigating whether reading into faces based on verbal information about behavior occurs at all, we were thus especially interested in the circumstances under which this RIF effect is more likely to (dis)appear. We attempted to visualize approximations of participants' mental representations using a data-driven reverse correlation (RC) methodology. We adopted a Bayesian inspired theoretical view on RIF, which led to two general predictions that we tested throughout the dissertation.

To introduce these general predictions, imagine that you are about to see the face of someone who previously robbed your best friend in the street. Our Bayesian inspired view assumes that the probability that this person's face ends up looking a certain way (say, untrustworthy) in your mind when you have seen the person's face depends both on (1) the probability that your brain would receive the sensory input it receives from this person's face when you look at it if there indeed were an untrustworthy looking face (termed likelihood), and (2) the prior probability of encountering an untrustworthy looking face in this situation (termed prior), which is likely influenced by the information that this person robbed your friend. In other words, two questions stand out: is the bottom-up visual input likely for an untrustworthy looking face and does the top-down prediction expect an untrustworthy looking face?

Based on these two paths of influence (bottom-up and top-down), the way in which verbal information about the person's criminal behavior could bias your mental representation of the person's seen face is by influencing the top-down expectation of encountering an untrustworthy looking face. This insight led to the first general prediction, which we termed the **Face Prior Hypothesis**: if verbal information about behavior is to bias the visual mental representation

of a seen face, it has to generate an expectation about facial appearance, called face prior. The second insight was that the relative strength of the expected facial appearance (face prior) and the input from the actual face (likelihood) determines their relative influence on your final mental representation after having seen the face. This insight led to the **Prior-Likelihood Balance Hypothesis**: verbal information is more (less) likely to bias the mentally represented facial appearance of a seen face if the face prior is relatively strong (weak) compared to input from the seen face. Thus, to understand under which circumstances verbal information about behavior is (not) likely to bias the visual mental representation of the seen face, we should find out which circumstances strengthen and weaken the face prior and likelihood, thereby altering their relative strength and hence their relative impact on the final mental representation of the seen face. Both general predictions are schematically represented in Figure 1.

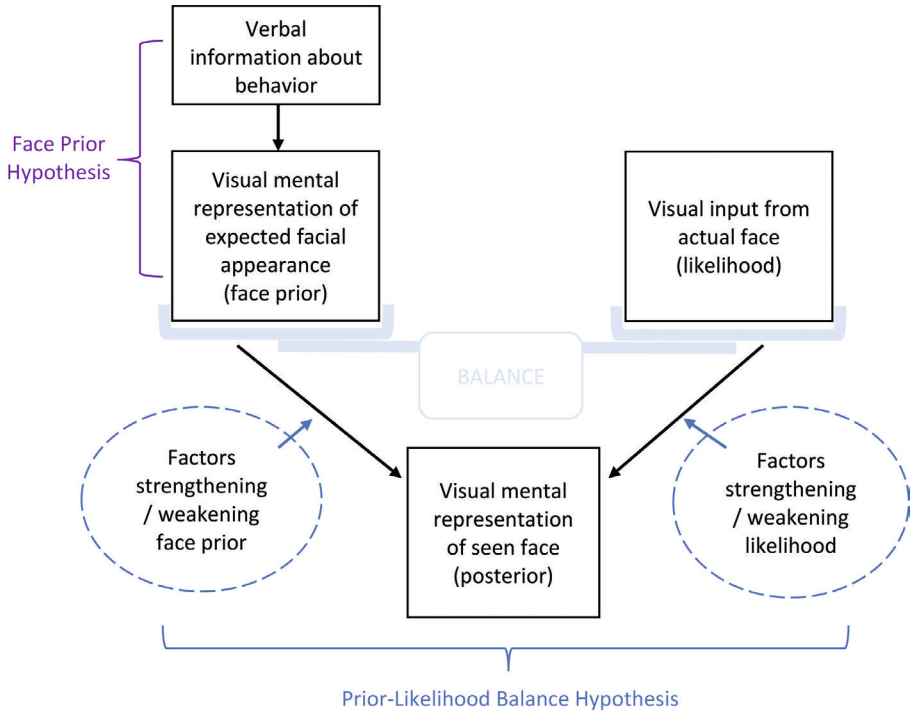


Figure 1. Schematic representation of the Face Prior Hypothesis and Prior-Likelihood Balance Hypothesis. The depicted balance scales represent the importance of the relative strength, suggesting that verbal information will only bias the mental representation of the seen face (posterior) if it leads to a face expectation (face prior) that is relatively strong compared to visual input of the actual face (likelihood).

Before investigating these two general predictions, we first laid relevant methodological groundwork in Chapter 2. The visualized approximations of participants' mental representations resulting from the RC task need to be scored on the concept of interest (in our case, trustworthiness appearance). Unfortunately, the currently available method is cost-inefficient when researchers want to conduct sequential hypothesis testing, as we planned to do. Therefore, we introduced and validated a more cost-efficient alternative to this scoring method and compared its results to that of the traditional method. Moreover, we demonstrated how researchers can create and validate their own version of this new method. Although our new method is more cost-efficient, comparisons to the traditional method showed that it is arguably less sensitive as well. Consequently, we proposed a combination of both scoring methods for the most efficient and optimal test when employing the RC task in combination

with sequential hypothesis testing. After this methodological contribution of Chapter 2, the remaining chapters investigated the main research question of this dissertation by focusing on the two general predictions derived from the Bayesian inspired theoretical view.

1. DOES VERBAL INFORMATION ABOUT BEHAVIOR GENERATE A VISUAL EXPECTATION ABOUT FACIAL APPEARANCE? (FACE PRIOR HYPOTHESIS)

We tested the first general prediction in Study 3.1. We manipulated verbal information about a target person's behavior with the aim to generate a trustworthiness impression (untrustworthy / trustworthy) about this person, confirmed in a manipulation check at the end of the study. After the verbal information and without ever showing the target's face, we measured participants' expectations of the person's facial appearance (i.e. their face priors) by having them visualize their mental representations of the expected face in a RC task and scoring these representations on trustworthiness appearance.

In line with scientific literature on visual stereotypes, the data strongly suggested that verbal information about behavior can indeed generate a visual face expectation. Participants' expectations of the target person's face appeared more trustworthy when he had performed positive opposed to negative behaviors. It is probable that this effect of behavioral information on the expected facial appearance runs through the formation of a person impression, such that the behavioral information creates an impression about the person's trustworthiness, which in turn influences the expected facial appearance.

2. DOES THE RELATIVE STRENGTH OF THE FACE PRIOR AND INPUT FROM THE SEEN FACE DETERMINE THE PROBABILITY OF A RIF EFFECT? (PRIOR-LIKELIHOOD BALANCE HYPOTHESIS)

We tested the second general prediction with five studies throughout Chapters 3 (Study 3.2-3.4) and 4 (Study 4.1-4.2). In all studies, we used the same experimental set-up as for the Face Prior Hypothesis in Study 3.1, except that we did present the face of the target person, namely after the behavioral

information and before the RC task. Importantly, we manipulated several factors intended to change the balance in strength between the face prior and likelihood across studies.

In Chapter 3, Study 3.2 encouraged participants to remember the face well (**memory instruction**) before presenting the face for 10 s. The data provided more evidence for the null, which stated that there was no RIF effect. In an attempt to weaken the likelihood, Study 3.3 omitted the memory instruction and limited the **face presentation duration** to 100 ms. Although the evidence shifted in the direction of a RIF effect, the effect was not yet convincing. Study 3.4 attempted to strengthen the face prior by having participants **mentally visualize their expectation of the face** before presenting the face for 100 ms. Under these conditions, we observed a subtle RIF effect such that the mental representation of the face looked slightly more trustworthy in the trustworthy opposed to untrustworthy condition. Thus, by increasingly shifting the balance in strength between the face prior and likelihood in favor of the prior with each study in Chapter 3, the evidence for a RIF effect became increasingly convincing.

In Chapter 4, we investigated whether a **time delay** of approximately 2 days between the face presentation and the RC task would increase the strength of the face prior in relation to the likelihood. This should result in a stronger RIF effect in a study including such a time delay (Study 4.1) than without such a time delay (Study 4.2). The idea was that it should be relatively easy for participants to remember their impression of the person as (un)trustworthy, which could be used to inform their face prior. The face prior should therefore remain quite stable over time. In comparison, it should be harder to remember the actual facial appearance of the target person, resulting in a weaker likelihood over time. Face presentation was set to 10 s for both studies with the intention to find more evidence for the null in Study 4.2 (without time delay), just like we had in Study 3.2. Because participants had been away from the study for approximately 2 days in Study 4.1, we asked them to refresh their memory by thinking back to their impression of the target person before starting the RC task in both studies. Both studies in Chapter 4 were thus identical, except for the time delay in Study 4.1. Surprisingly, we found evidence for a RIF effect not only in Study 4.1, but also in Study 4.2. Comparisons of Study 4.2 and Study 3.2 suggested that the most likely cause for this unexpected RIF effect in Study 4.2 was the instruction

to **activate the person impression**, which probably strengthened the face prior. Although the combined data of Study 4.1 and 4.2 carefully suggested that a time delay may lead to a stronger RIF effect when the target's actual face appears neutral on trustworthiness, more research is needed to provide convincing evidence for or against this suggestion.

Taken together, the patterns of results across Chapter 3 and 4 provide support for the Prior-Likelihood Balance Hypothesis. Whenever we strengthened the prior in relation to the likelihood, the RIF effect either remained quite similar or became more pronounced. Importantly, the RIF effect never became weaker under these circumstances, which would have contradicted the Prior-Likelihood Balance Hypothesis. For RIF to occur, it thus seems important that verbal information generates a face prior that is strong enough to compete with the sensory input from the actual face. In other words, the data support the importance of the balance scales depicted in Figure 1.

Regarding the specific circumstances that impact this balance (depicted in the blue circles in Figure 1), especially the activation of a person impression appeared successful, as a subtle RIF effect emerged convincingly in studies including such an activation (Study 3.4, and 4.1-4.2) but not in studies without such an activation (Study 3.3 and 3.2). This impression could be either of the person in general (Study 4.1-4.2) or specifically about the expected facial appearance (Study 3.4). The other factors investigated appeared to have far less impact on the emergence and strength of RIF, at least for the specific manipulations that we investigated. However, it remains interesting to see how a memory instruction's workings against the emergence of RIF would hold when combined with an active person impression. Future studies could clarify this by manipulating the presence of a memory instruction while always ensuring an active person impression. Moreover, future studies could investigate whether a longer time delay than 2 days would lead to a stronger RIF effect or whether the currently chosen time delay would increase the RIF effect when the person impression is not activated. The five factors that we investigated across studies can be filled in the blue circles of Figure 1 together with an indication of their probable successfulness, as shown in Figure 2.

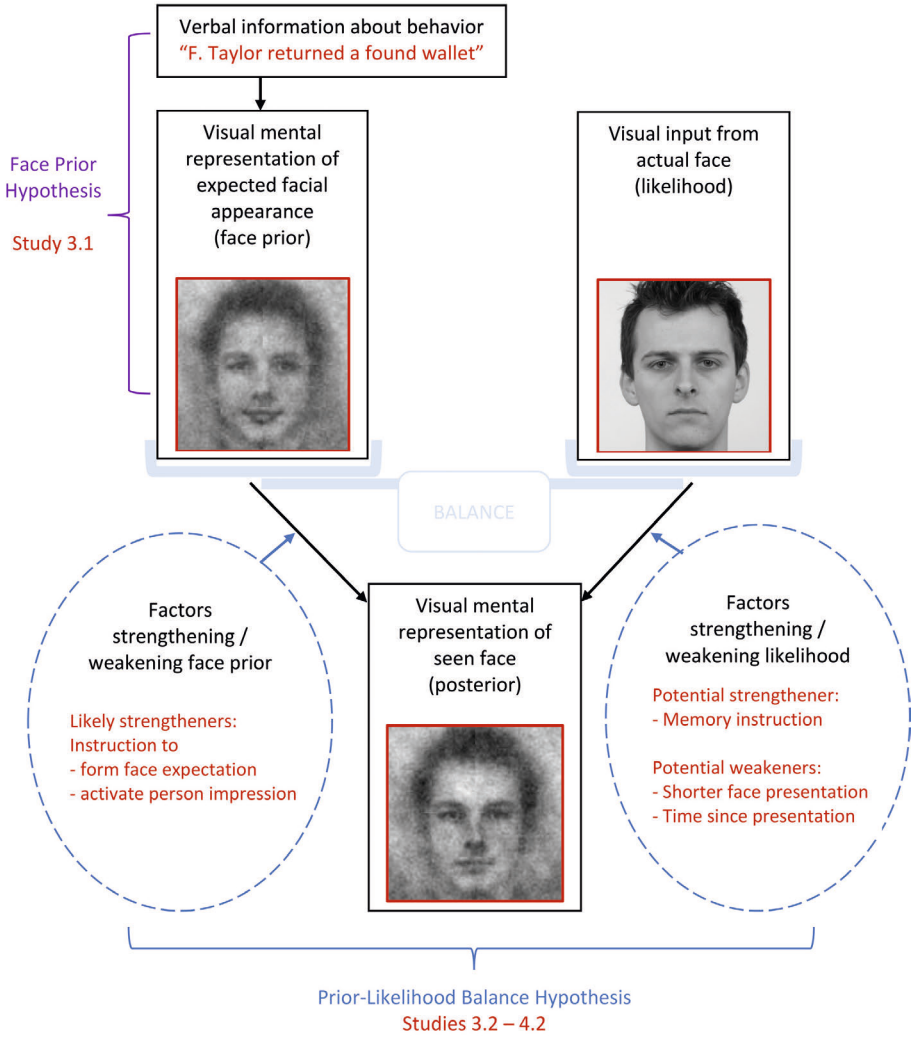


Figure 2. Schematic representation of the two general predictions with examples of manipulations and measurements from our studies added in red (likelihood and posterior examples taken from Study 3.4). Due to space restrictions, only examples from the 'trustworthy' experimental condition are portrayed. The varying circumstances which were aimed at manipulating the relative strength of the expected facial appearance (face prior) and input from the actual face (likelihood) are added in the blue circles with indications of their successfulness.

In sum, the most important specific circumstance for the emergence of RIF seems to be an active person impression (informed by the verbal information) when mentally representing the seen face. This is consistent with scientific literature showing that attitudes and stereotypes should be accessible in order

to influence evaluations. Thus, even if the verbal information generates a person impression that is associated with certain facial appearances, it will bias the mental representation of the seen face only if this impression is indeed accessible (i.e. active) at the time of mentally representing the seen face. Interestingly, it is possible that one also should not have the goal to remember the face very accurately, but further research is needed to clarify this.

CONCLUSION

The present dissertation provides both a methodological and theoretical contribution to the scientific field of social face perception. Using a data-driven, innovated RC methodology and a Bayesian inspired theoretical approach, the present dissertation suggests that verbal information about behavior can subtly bias mental representations of a seen face in line with the behavioral information, but only when it generates a face prior that is relatively strong compared to input from the actual face. The most important factor for this to be the case seems to be the accessibility of a person impression of the target (which likely informs the face prior). In other words, for RIF to occur, it appears that the verbal information should generate a person impression that can be associated with certain facial features and that is active while mentally representing the seen face.

Interestingly, verbal information can thus indeed influence visual mental representations of seen faces, substantiating the idea that our mental representations of faces can be informed by a combination of top-down expectations and bottom-up sensory input. Hopefully, these findings help to raise awareness about the circumstances that may create biases in our social impressions. Now that we have made a start in demonstrating and understanding the context dependency of RIF, we encourage researchers to replicate our findings (using both similar and different stimuli and manipulations) and to reveal even more moderating factors as well as potential social consequences of this intriguing bias in social face perception. Hopefully, as insights into RIF continue to accumulate, people can take these insights into account when making socially consequential decisions about others, for instance when describing a perpetrator's face as an eyewitness or when determining someone's fate in politics, the courtroom, or on the financial, labor, or housing market.

Nederlandse samenvatting (Dutch summary)

Gezichten spelen een belangrijke rol in het sociale leven van mensen. Niet alleen krijgen gezichten de meeste aandacht in alledaagse sociale interacties, ze hebben ook een diepgaande invloed op hoe de waarnemer denkt, zich voelt, en zich gedraagt tegenover de persoon achter het gezicht, waardoor sociale beslissingen worden beïnvloed die gevolgen hebben die verder gaan dan die specifieke sociale interactie. Wetenschappelijk onderzoek naar sociale gezichtswaarneming heeft bijvoorbeeld aangetoond dat iemands gezichtsverschijning (d.w.z. hoe iemands gezicht eruitziet) invloed kan hebben op dermate belangrijke beslissingen als het al dan niet stemmen op iemand in de politiek, het aannemen van iemand voor een baan, het vertrouwen in iemand op financieel vlak, of zelfs het ter dood veroordelen van een beklaagde.

Echter, het cognitieve systeem van de mens werkt niet als een passief opnameapparaat, maar construeert juist op een actieve manier een eigen mentale representatie van de realiteit, gebaseerd op de zintuiglijke input die het ontvangt van de wereld "daarbuiten" en de beschikbare kennis over de wereld. Dit roept de vraag op hoe accuraat mensen de gezichten van anderen mentaal representeren, met name wanneer zij al enige kennis over de ander hebben. Kunnen mentale representaties van gezichten, naast het beïnvloeden van sociale evaluaties, zelf ook beïnvloed worden door sociale informatie? Als dat zo is, lezen mensen niet alleen informatie *van* gezichten *af*, maar lezen ze het er ook *in*.

Wetenschappelijk onderzoek heeft aangetoond dat sociale visuele signalen, zoals haardracht, kleding, gender en etniciteit, de sociale categorisatie van een gezicht inderdaad kunnen beïnvloeden, bijvoorbeeld op etniciteit, gender, en emotionele expressie. Maar hoe zit dat met sociale signalen die niet samen met het gezicht aanwezig zijn in het visuele veld? Zou sociale kennis gebaseerd op verbale informatie ook invloed kunnen hebben op de mentale representatie van een gezicht, zelfs als men dat gezicht gezien heeft? Als dat zo is, zou dat het theoretische idee onderbouwen dat mentale representaties van gezichten beïnvloed kunnen worden door zowel zintuiglijke input van het gezicht dat "daarbuiten" is, alsook door andere sociale overtuigingen die de waarnemer mogelijk over de waargenomen persoon heeft. Bovendien zou het impliceren dat mentale representaties van iemands gezicht niet alleen beïnvloed kunnen worden door die persoon zelf en het selecte groepje dat zijn/haar visuele

verschijning kan veranderen (bijvoorbeeld de kapper), maar in feite door iedereen die enigszins geloofwaardige informatie over die persoon kan delen (bijvoorbeeld via roddels).

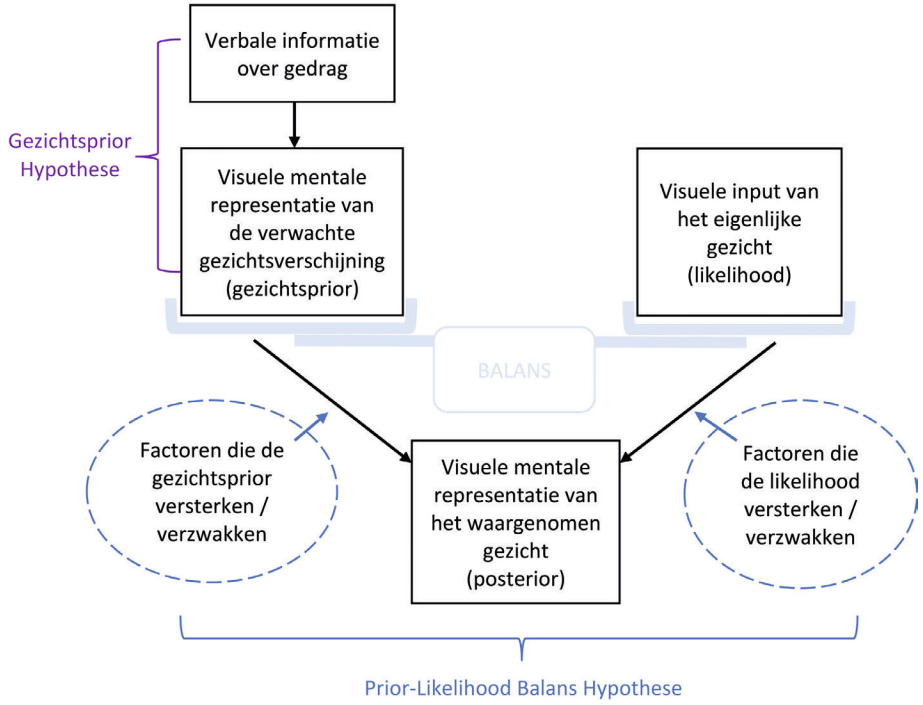
Het is daarom zowel theoretisch als maatschappelijk relevant om te begrijpen of, en onder welke omstandigheden verbale informatie over een persoon invloed kan hebben op waarnemers' mentale representaties van het gezicht van die persoon, zelfs wanneer ze dat gezicht gezien hebben. Zulke inzichten kunnen ons theoretische begrip vergroten over de mate waarin en de omstandigheden waaronder onze cognitieve systemen gebruiken maken van beschikbare sociale informatie die niet visueel aanwezig is om de mentale representatie van iemands gezicht te vormen. Bovendien kan een dergelijk begrip helpen om maatschappelijk bewustzijn te creëren over de omstandigheden waaronder onze mentale representaties van andermans gezichten waarschijnlijk gekleurd zijn.

Het doel van dit proefschrift was om meer inzicht te krijgen in de omstandigheden waaronder verbale informatie over iemands gedrag waarschijnlijk (g)een vertekenend effect heeft op de visuele mentale representatie van diens waargenomen gezicht, ook wel 'reading into faces' (RIF) genoemd. Naast de vraag of RIF gebaseerd op verbale informatie over gedrag überhaupt voorkomt, waren we dus vooral geïnteresseerd in de omstandigheden waaronder dit RIF effect waarschijnlijker is om (niet) op te treden. We trachtten benaderingen van deelnemers' mentale representaties te visualiseren met behulp van een data-gedreven 'reverse correlation' (RC) methodologie. Daarnaast hanteerden we een Bayesiaans geïnspireerde theoretische visie op RIF, welke resulteerde in twee algemene voorspellingen die we in het huidige proefschrift testten.

Ter introductie van deze algemene voorspellingen: stel je voor dat je op het punt staat het gezicht te zien van iemand die eerder jouw beste vriend(in) op straat heeft beroofd. Onze Bayesiaans geïnspireerde zienswijze gaat ervan uit dat de waarschijnlijkheid dat het gezicht er op een bepaalde manier (bijvoorbeeld onbetrouwbaar) uitziet in jouw hoofd wanneer je het gezicht van die persoon hebt gezien, afhangt van de volgende twee zaken. (1) De waarschijnlijkheid dat jouw hersenen de zintuiglijke input die zij van het gezicht van deze persoon ontvangen wanneer je ernaar kijkt, zouden ontvangen als er inderdaad een onbetrouwbaar uitziend gezicht zou zijn ('likelihood' genoemd). (2) De

voorafgaande waarschijnlijkheid dat je in deze situatie een onbetrouwbaar uitziend gezicht tegenkomt ('prior' genoemd), welke mogelijk beïnvloed is door de informatie dat deze persoon jouw vriend(in) beroofd heeft. Met andere woorden, is de bottom-up visuele input waarschijnlijk voor een onbetrouwbaar uitziend gezicht en verwacht de top-down voorspelling een onbetrouwbaar uitziend gezicht?

Gebaseerd op deze twee beïnvloedingspaden (bottom-up en top-down), blijkt dat verbale informatie over het criminele gedrag van de persoon jouw mentale representatie van diens waargenomen gezicht kan beïnvloeden door middel van het beïnvloeden van de top-down verwachting om een onbetrouwbaar uitziend gezicht tegen te komen. Dit inzicht leidde tot de eerste algemene voorspelling, die we de **Gezichtsprior Hypothese** noemden: als verbale informatie over gedrag de visuele mentale representatie van een waargenomen gezicht wil kunnen beïnvloeden, moet de verbale informatie een verwachting over het gezichtsuiterlijk genereren, de zogenaamde gezichtsprior. Het tweede inzicht was dat de relatieve sterkte van de verwachte gezichtsverschijning (gezichtsprior) en de input van het werkelijke gezicht (likelihood) hun relatieve invloed bepaalt op jouw uiteindelijke mentale representatie nadat je het gezicht hebt gezien. Dit inzicht leidde tot de **Prior-Likelihood Balans Hypothese**: het is meer (minder) waarschijnlijk dat verbale informatie de mentaal gepresenteerde gezichtsverschijning van een waargenomen gezicht zal vertekenen als de gezichtsprior relatief sterk (zwak) is in vergelijking met de input van het waargenomen gezicht. Om te begrijpen onder welke omstandigheden verbale informatie over gedrag de visuele mentale representatie van een waargenomen gezicht waarschijnlijk (niet) zal vertekenen, moeten we daarom uitzoeken welke omstandigheden de gezichtsprior en likelihood versterken en verzwakken, waardoor hun relatieve sterkte verandert, en daarmee hun relatieve impact op de uiteindelijke mentale representatie van het waargenomen gezicht. Beide algemene voorspellingen zijn schematisch weergegeven in Figuur 1.



Figuur 1. Schematische voorstelling van de Gezichtsprior Hypothese en de Prior-Likelihood Balans Hypothese. De afgebeelde weegschaal geeft het belang van de relatieve sterkte weer, wat suggereert dat verbale informatie alleen de mentale representatie van het waargenomen gezicht (posterior) zal vertekenen als het leidt tot een gezichtsverwachting (gezichtsprior) die relatief sterk is in vergelijking met visuele input van het eigenlijke gezicht (likelihood).

Alvorens deze twee algemene voorspellingen te onderzoeken, hebben we in Hoofdstuk 2 eerst een belangrijke methodologische basis gelegd. De RC taak levert gevisualiseerde benaderingen van deelnemers' mentale representaties op en die moeten gescoord worden op het concept waarin de onderzoekers geïnteresseerd zijn (in ons geval was dat hoe betrouwbaar de gezichten oogden). Helaas is de huidige beschikbare scoringsmethode kosten-inefficiënt wanneer onderzoekers sequentiële hypothesetests willen uitvoeren, zoals wij van plan waren te doen. Daarom introduceerden en valideerden we een kosten-efficiënter alternatief voor deze scoringsmethode en vergeleken de resultaten ervan met die van de traditionele methode. Daarnaast hebben we gedemonstreerd hoe onderzoekers hun eigen versie van deze nieuwe methode kunnen creëren en valideren. Hoewel onze nieuwe methode kosten-efficiënter is, impliceerden vergelijkingen met de traditionele methode dat onze methode ook minder

sensitief is. Derhalve hebben we een combinatie van beide scoringsmethoden voorgesteld voor de meest efficiënte en optimale test bij het gebruik van de RC taak in combinatie met sequentiële hypothesetoetsing. Na deze methodologische bijdrage van Hoofdstuk 2, werd in de resterende hoofdstukken de hoofdonderzoeksvraag van dit proefschrift onderzocht door te focussen op de twee algemene voorspellingen die werden afgeleid uit de Bayesiaans geïnspireerde theoretische visie.

1. GENEREERT VERBALE INFORMATIE OVER GEDRAG EEN VISUELE VERWACHTING OVER GEZICHTSVERSCHIJNING? (GEZICHTSPRIOR HYPOTHESE)

We hebben de eerste algemene voorspelling onderzocht in Studie 3.1. We manipuleerden verbale informatie over het gedrag van een persoon met het doel een betrouwbaarheidsindruk (onbetrouwbaar / betrouwbaar) over deze persoon te genereren, wat werd bevestigd in een manipulatiecheck aan het einde van de studie. Na de verbale informatie en zonder ooit het gezicht van de persoon te hebben getoond, maten we deelnemers' verwachtingen over de gezichtsverschijning van de persoon (d.w.z. hun gezichtspriors) door hen hun mentale representaties van het verwachte gezicht te laten visualiseren in een RC taak en deze representaties te scoren op hoe betrouwbaar deze oogden.

In lijn met wetenschappelijke literatuur over visuele stereotypen, suggereerden de data sterk dat verbale informatie over gedrag inderdaad een visuele gezichtsverwachting kan opwekken. Deelnemers' verwachtingen van het gezicht van de persoon zagen er betrouwbaarder uit wanneer de gedragingen van de persoon vooral positief in tegenstelling tot negatief waren geweest. Het is waarschijnlijk dat dit effect van gedragsinformatie op de verwachte gezichtsverschijning loopt via de vorming van een persoonsindruk, zodanig dat de gedragsinformatie een indruk wekt over de betrouwbaarheid van de persoon, welke vervolgens de verwachte gezichtsverschijning beïnvloedt.

2. BEPAALT DE RELATIEVE STERKTE VAN DE GEZICHTSPRIOR EN DE INPUT VAN HET WAARGENOMEN GEZICHT DE WAARSCHIJNLIJKHEID VAN EEN RIF EFFECT? (PRIOR-LIKELIHOOD BALANS HYPOTHESE)

We hebben de tweede algemene voorspelling onderzocht aan de hand van vijf studies in Hoofdstukken 3 (Studie 3.2-3.4) en 4 (Studie 4.1-4.2). In alle studies hebben we dezelfde experimentele opzet gebruikt als voor de Gezichtsprior Hypothese in Studie 3.1, behalve dat we het gezicht van de persoon nu wel presenteerden, namelijk na de gedragsinformatie en voor de RC taak. Van belang is dat we verschillende factoren hebben gemanipuleerd met het doel om de balans in sterkte tussen de gezichtsprior en likelihood te veranderen tussen studies.

In Hoofdstuk 3 spoorden we deelnemers in Studie 3.2 aan om het gezicht goed te onthouden (**geheugeninstructie**) voordat we het gezicht voor 10 s presenteerden. De data leverden meer bewijs voor het nul model dat er geen RIF effect was. In een poging om de likelihood te verzwakken, lieten we in Studie 3.3 de geheugeninstructie weg en beperkten we de **duur van de gezichtspresentatie** tot 100 ms. Hoewel het bewijs wat verschoof in de richting van een RIF effect, was het effect nog niet overtuigend. In Studie 3.4 trachtten we de gezichtsprior te versterken door deelnemers **hun verwachting van het gezicht mentaal te laten visualiseren** voordat we het gezicht voor 100 ms presenteerden. Onder deze omstandigheden observeerden we een subtiel RIF effect, zodanig dat de mentale representatie van het waargenomen gezicht er iets betrouwbaarder uitzag in de betrouwbaarheids- versus onbetrouwbaarheidsconditie. Kortom, door met elke opeenvolgende studie in Hoofdstuk 3 de balans in sterkte tussen de gezichtsprior en likelihood steeds iets te verschuiven in het voordeel van de prior, werd het bewijs voor een RIF effect steeds iets overtuigender.

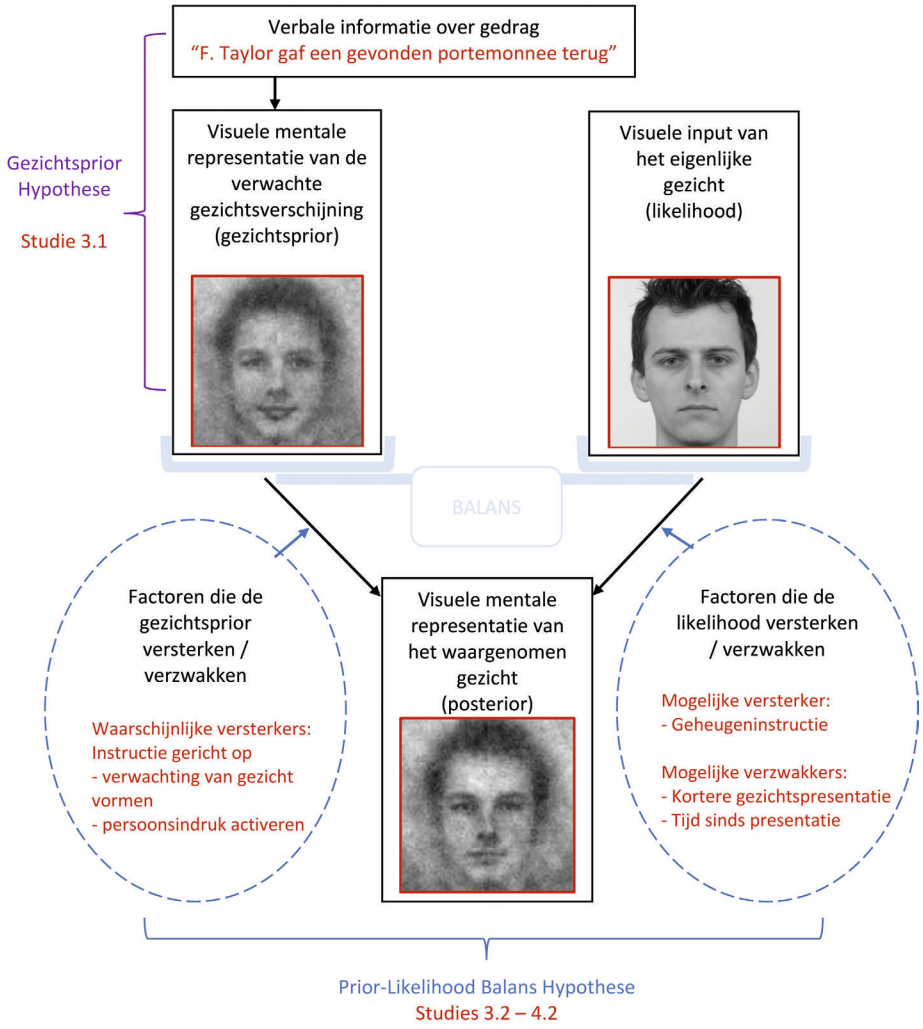
In Hoofdstuk 4 hebben we onderzocht of een **tijdvertraging** van ongeveer 2 dagen tussen de presentatie van het gezicht en de RC taak de sterkte van de gezichtsprior ten opzichte van de likelihood zou verhogen. Dit zou moeten resulteren in een sterker RIF effect in een studie met zo'n vertraging (Studie 4.1) dan een studie zonder zo'n vertraging (Studie 4.2). Het idee was dat het relatief gemakkelijk voor deelnemers zou moeten zijn om hun indruk van de

persoon als (on)betrouwbaar te onthouden, wat gebruikt zou kunnen worden om hun gezichtsprior te vormen. De gezichtsprior zou daarom vrij stabiel moeten blijven in de tijd. In vergelijking zou het moeilijker moeten zijn om het eigenlijke gezicht van de persoon te onthouden, wat zou moeten resulteren in een zwakkere likelihood na verloop van tijd. De presentatieduur van het gezicht was voor beide studies ingesteld op 10 s, met de bedoeling om meer bewijs voor het nul model te vinden in Studie 4.2 (zonder tijdvertraging), net zoals we dat in Studie 3.2 hadden gevonden. Omdat deelnemers in Studie 4.1 ongeveer 2 dagen weg waren geweest van de studie, vroegen we deelnemers in beide studies hun geheugen op te frissen door terug te denken aan hun indruk van de persoon voordat ze aan de RC taak begonnen. Beide studies in Hoofdstuk 4 waren dus identiek, met uitzondering van de tijdvertraging in Studie 4.1. Verrassend genoeg vonden we niet alleen bewijs voor een RIF effect in Studie 4.1, maar ook in Studie 4.2. Vergelijkingen tussen Studie 4.2 en Studie 3.2 suggereerden dat de meest waarschijnlijke oorzaak van dit onverwachte RIF effect in Studie 4.2 de instructie was om **de persoonsindruk te activeren**, wat waarschijnlijk de gezichtsprior versterkte. Hoewel de gecombineerde data van Studie 4.1 en 4.2 voorzichtig suggereerden dat een tijdvertraging zou kunnen leiden tot een sterker RIF effect wanneer het werkelijke gezicht vrij neutraal oogt op betrouwbaarheid, is er meer onderzoek nodig om overtuigend bewijs voor of tegen deze suggestie te leveren.

Samengenomen bieden de resultaten van Hoofdstuk 3 en 4 steun voor de Prior-Likelihood Balans Hypothese. Telkens wanneer we de prior versterkten in verhouding tot de likelihood, bleef het RIF effect ofwel vrijwel gelijk ofwel werd het uitgesprokener. Belangrijk is dat het RIF effect onder deze omstandigheden nooit zwakker werd, wat de Prior-Likelihood Balans Hypothese zou hebben tegengesproken. Om RIF te laten optreden, lijkt het dus belangrijk dat verbale informatie een gezichtsprior genereert die sterk genoeg is om te concurreren met de zintuiglijke input van het werkelijke gezicht. Met andere woorden, de data ondersteunen het belang van de weegschaal afgebeeld in Figuur 1.

Wat betreft de specifieke omstandigheden die het evenwicht tussen deze schalen beïnvloeden (afgebeeld in de blauwe cirkels in Figuur 1), bleek met name de activering van een persoonsindruk belangrijk, aangezien een subtiel RIF effect overtuigend optrad in studies inclusief een dergelijke activering (Studie 3.4, en

4.1-4.2), maar niet in studies exclusief een dergelijke activering (Studie 3.3 en 3.2). Deze indruk kon zowel betrekking hebben op de persoon in het algemeen (Studie 4.1-4.2) als specifiek op de verwachte gezichtsverschijning (Studie 3.4). De andere onderzochte factoren bleken aanzienlijk minder invloed te hebben op het ontstaan en de sterkte van RIF, althans voor de specifieke manipulaties die wij onderzochten. Het blijft echter interessant om te zien hoe de werking van een geheugeninstructie tegen het ontstaan van RIF stand zou houden wanneer deze wordt gecombineerd met een actieve persoonsindruk. Toekomstig onderzoek zou dit kunnen verduidelijken door de aanwezigheid van een geheugeninstructie te manipuleren terwijl er altijd een actieve persoonsindruk is. Daarnaast zouden toekomstige studies kunnen onderzoeken of een langere tijdvertraging dan 2 dagen tot een sterker RIF effect zou kunnen leiden, of dat de huidige gekozen tijdvertraging voldoende zou zijn wanneer de persoonsindruk niet speciaal geactiveerd is. De vijf factoren die we over studies heen hebben onderzocht, kunnen samen met een indicatie van de waarschijnlijkheid van hun succes worden ingevuld in de blauwe cirkels van Figuur 1, zoals hieronder weergegeven in Figuur 2.



Figuur 2. Schematische representatie van de twee algemene voorspellingen met voorbeelden van de manipulaties en metingen uit onze studies toegevoegd in rood (likelihood en posterior voorbeelden komen uit Studie 3.4). Wegens ruimtegebrek zijn alleen voorbeelden uit de 'betrouwbare' experimentele conditie afgebeeld. De verschillende omstandigheden die erop gericht waren om de relatieve sterkte van de verwachte gezichtsverschijning (gezichtsprior) en input van het werkelijke gezicht (likelihood) te manipuleren, zijn samen met een indicatie van hun succes toegevoegd in de blauwe cirkels.

Kortom, de belangrijkste specifieke omstandigheid voor het ontstaan van RIF lijkt te zijn dat de waarnemer een actieve persoonsindruk (geïnformeerd door de verbale informatie) moet hebben tijdens het mentaal representeren van het waargenomen gezicht. Dit komt overeen met wetenschappelijke literatuur die

aantoont dat attitudes en stereotypen toegankelijk moeten zijn om evaluaties te kunnen beïnvloeden. Dus zelfs als de verbale informatie een persoonsindruk genereert die geassocieerd is met bepaalde gezichtskenmerken, zal dit alleen leiden tot een vertekende mentale representatie van het waargenomen gezicht als deze indruk inderdaad toegankelijk (d.w.z. actief) is op het moment van het mentaal representeren van het waargenomen gezicht. Mogelijk is het ook van belang dat de waarnemer niet het doel heeft om het gezicht heel accuraat te onthouden, maar meer onderzoek is nodig om dit te verduidelijken.

CONCLUSIE

Het huidige proefschrift levert zowel een methodologische als theoretische bijdrage aan het wetenschappelijke onderzoeksveld van sociale gezichtswaarneming. Gebruikmakend van een data-gedreven, geïnnoveerde RC methodologie en een Bayesiaans geïnspireerde theoretische benadering, suggereert het proefschrift dat verbale informatie over gedrag op subtiele wijze mentale representaties van een waargenomen gezicht in lijn met de gedragsinformatie kan vertekenen, echter alleen wanneer deze informatie een gezichtsprior genereert die relatief sterk is ten opzichte van de input van het werkelijke gezicht. De belangrijkste factor om dit te bereiken lijkt de toegankelijkheid van een persoonsindruk te zijn (die waarschijnlijk de gezichtsprior beïnvloedt). Met andere woorden, om RIF te laten plaatsvinden, lijkt het erop dat de verbale informatie een persoonsindruk moet genereren die geassocieerd is met bepaalde gezichtskenmerken en die actief is tijdens het mentaal representeren van het waargenomen gezicht.

Interessant is dat verbale informatie dus wel degelijk visuele mentale representaties van waargenomen gezichten kan beïnvloeden. Dit onderbouwt het idee dat onze mentale representaties van gezichten gevormd kunnen worden door een combinatie van top-down verwachtingen en bottom-up zintuiglijke input. Hopelijk helpen deze bevindingen om meer bewustwording te creëren rondom de omstandigheden die vertekeningen in onze sociale indrukken kunnen veroorzaken. Nu we een begin hebben gemaakt in het aantonen en begrijpen van de contextafhankelijkheid van RIF, nodigen we onderzoekers uit om onze bevindingen te repliceren (met zowel soortgelijke als verschillende stimuli en manipulaties) en om nog meer modererende factoren en potentiële

sociale gevolgen van deze intrigerende bias in sociale gezichtswaarneming aan het licht te brengen. Als inzichten in RIF zich blijven opstapelen, kunnen mensen hopelijk rekening houden met deze inzichten wanneer ze besluiten over anderen nemen die sociale gevolgen met zich meedragen, zoals het omschrijven van het gezicht van een dader als ooggetuige, of het bepalen van iemands lot in de politiek, de rechtszaal, of op de financiële, arbeids-, of huizenmarkt.

Dankwoord (Acknowledgements)

Het is een hobbelige rit geweest om tot de voltooiing van dit proefschrift te komen. Een rit die ik nooit in mijn eentje had kunnen – noch willen – volbrengen. Wat een geweldige mensen hebben er met mij mee gehobbeld en wat hebben ze de rit enorm veel waardevoller, plezieriger, leerzamer, en mooier gemaakt. Lekker mee verheugen op momenten dat de vaart er goed in zat, mee fantaseren over de dingen die we onderweg tegenkwamen en wat dat dan betekende, helpen duwen wanneer ik in de modder vast zat, accepteren hoe ik tot drie keer toe van route veranderde, samen balen op het moment dat mijn vervoersmiddel afsloeg, eindeloos veel geduld en steun bieden toen die ook nog eens heel lang in de garage gerepareerd moest worden, toejuichen als er weer een kilometertje bij kwam, af en toe wat bijsturen, super leuke afleidingen bieden, lekker lang discussiëren over kleine details in het reisverslag, en zo nu en dan helpen om de boel een beetje te relativeren. En dat alles terwijl ik altijd alle ruimte heb gevoeld om zelf achter het stuur te mogen blijven zitten. Mijn dank is groot aan eenieder die de afgelopen jaren op deze reis heeft bijgedragen aan mijn ontwikkeling als wetenschapper en als persoon, hoe groot of klein ook. Een aantal mensen wil ik hier in het bijzonder bedanken.

Ten eerste mijn zeer getalenteerde team van begeleiders: Daniël, Ron, en Rob. Ik had jullie alle drie al hoog in aanzien staan tijdens mijn bachelor en master en ik kan oprecht zeggen dat mijn waardering voor jullie tijdens mijn PhD project alleen maar is gegroeid. Ik vind jullie stuk voor stuk zowel voortreffelijke wetenschappers als mooie persoonlijkheden. Samenwerken met jullie vond ik een feestje. Jullie kritische vragen hielpen mij om een betere wetenschapper te worden. En jullie humor en inlevingsvermogen zorgden ervoor dat ik altijd met een positief gevoel uit onze afspraken kwam. Ik prijs mezelf gelukkig dat ik zo lang met jullie heb mogen samenwerken. Ja, want lang was het zeker. En dat terwijl jullie allemaal binnen korte tijd na het aansluiten aan mijn PhD project zelf flinke carrièrestappen zetten en daardoor (gedeeltelijk) andere verantwoordelijkheden kregen. Hoewel ik mezelf graag voorhoud dat een samenwerking met mij altijd op magische wijze leidt tot gegarandeerd carrièresucces, reflecteert het natuurlijk alleen maar hoe supercompetent jullie zijn. Bedankt dat jullie al die tijd betrokken zijn gebleven bij mijn PhD project. Een paar woorden voor ieder van jullie persoonlijk:

Daniël, de wijze en genietende verhalenverteller. Al vanaf het eerste jaar van mijn bachelor heb jij psychologie leuk en levendig voor mij gemaakt met je entertainende AIP colleges. En wat blijkt, niet alleen tijdens colleges, maar ook tijdens onderzoekafspraken maak je altijd even tijd om een leuk verhaal te delen. Eigenlijk heb je het ontzettend druk, maar toch kom je altijd rustig over. Je straalt plezier uit in wat je doet en dat werkt aanstekelijk. Daarnaast heb je een subliem oog voor detail, iets wat ik goed kan waarderen. Ik denk met plezier terug aan lange discussies die we hebben gevoerd over enkele woordjes. Bovendien heb je oog voor de ander. Je geeft om het welzijn van anderen, wat ik met name heb ervaren toen ik ziek werd. Het feit dat je me eerder probeerde af te remmen dan aan te sporen was voor mij een teken dat je mij goed kende en mij zag, want dat was precies wat ik op dat moment nodig had. Bedankt voor al je steun, je wijsheid, je vertrouwen, en je aanstekelijke optimisme.

Ron, de supersociale supernerd. Wat heb ik veel van jou geleerd. Al in de bachelor zorgde je voor hele coole academische ervaringen toen Lianne en ik jou en Alexander Todorov mochten opzoeken aan Princeton University. Ik heb jou altijd prettig benaderbaar gevonden en heb altijd het gevoel gehad dat ik met alle twijfels over de academische wereld bij je terecht kan. Het was dan ook na een gesprek met jou tijdens mijn master dat ik de knoop had doorgehakt om een PhD project te gaan doen. Dankzij jouw expertise en heldere manier van uitleggen heb ik bovendien hele boeiende onderzoeksmethoden kunnen gebruiken en kan ik ze nog goed begrijpen ook. Jouw scherpe vragen hebben me ook enorm geholpen om kritisch te kijken naar wat de resultaten nou betekenen. Daarnaast heb je het ook nog eens mogelijk gemaakt dat Béla en ik konden gaan samenwonen op een heel leuk plekje in Nijmegen. Bedankt voor je luisterend oor, je relativeringsvermogen, je grapjes, en je waardevolle wetenschaps-, levens-, en Bali-adviezen.

Rob, de vrolijke en intelligente redder in nood. Toen geen enkele begeleider meer binnen onze afdeling werkte, sloot jij je aan bij ons project. Held! Wat was en ben ik daar blij mee. Ik vond dat ik het altijd al goed met jou kon vinden, dus toen Daniël en Ron vertelden dat ze jou gevraagd hadden, was ik meteen enthousiast. Ik vond het een plezier om met jou samen te mogen werken. Je straalt vriendelijkheid en enthousiasme uit en bent goed in wat je doet. En je houdt rekening met iedereen, waardoor ik mij echt gezien voelde. Dankzij

jouw grondige feedback zijn de hoofdstukken in dit proefschrift zoveel beter geworden. Dan dacht ik in eerste instantie dat het al best goed was, maar dan zag ik tijdens het herschrijven eigenlijk altijd dat het van jouw feedback nog veel beter werd. Bedankt voor je onmisbare inzet, je steun en respect, je rake feedback, en je aanstekelijke vrolijkheid en enthousiasme.

Ook wil ik Ad graag bedanken. Hoewel je niet betrokken bent geweest bij het onderzoek in dit proefschrift, heb je wel aan het begin van mijn PhD project gestaan. Ik vond het een voorrecht dat ik even heb mogen profiteren van jouw enorme wijsheid en vriendelijke aanwezigheid. Ik zie je nog met pretogen in de stoel in Daniëls kantoor zitten. Jouw afscheidsmail aan mij, waarin je jouw vertrouwen in mij uitsprak, raakte mij diep. Bedankt Ad!

Loek, bedankt voor het warme welkom in Utrecht en voor de leuke samenwerking op Hoofdstuk 3. Bedankt voor je openheid (zowel in de wetenschap als persoonlijk) en voor de boeiende gesprekken.

Gesa, mijn lieve vriendinnetje, roomie, en paranimf. Mijn PhD was half niet zo leuk geweest zonder jou. Wat had ik gemoeten zonder jouw liefdevolle knuffels, creatieve tekeningetjes, diepgaande gesprekken en adviezen, en onmisbare schoudermassages. Ik hou van hoe prettig gestoord wij samen kunnen zijn. Gedurende onze tijd als roomies hebben we onze eigen communicatie ontwikkeld die zich niet beperkt tot de woorden in het woordenboek (en die af en toe zelfs uit enkel gekke geluidjes bestaat). Onze spontane uitspattingen aan kantoorfitness, wandelingetjes in de buitenlucht, meditatiemomentjes, te gekke dansjes, en gekleurde zitballen hebben mij enorm geholpen om de kantoordagen achter de computer vol te kunnen houden. Maar ook buiten het kantoor heb je me zoveel steun en plezier gegeven. Bedankt voor al die liefde. Ik hoop dat we elkaar nog lang zullen blijven opzoeken.

André, mijn academische “broer” en paranimf. Jouw kalmte, vriendelijkheid, en enorme intelligentie maken jou tot een gesprekspartner waar ik uren mee kan blijven praten. En dan hebben we ook nog eens veel van dezelfde interesses! Bedankt voor alle stimulerende en leuke gesprekken over wetenschappelijke theorieën, dansstijlen, mooie reizen, en gezondheidskwesties. De EASP summer school in Lissabon was een stuk leuker met jou erbij. Ik heb je dan ook gemist

op de afdeling na je vertrek en ben blij en vereerd dat je als paranimf weer naast me komt staan.

Bernice. Having you as my roomie made each office day more fun. Thank you for the countless times that you listened to my struggles and rejoiced in my progressions. You radiate so much love, warmth, and understanding. I felt I could tell you everything. And I could always count on you to do or say something funny and put that smile right back on my face. I cherish my memories of our Diehard Office outings (with Gesa) to the zoo, the little muffin shop, the Vierdaagse Feesten, and of course of singing along to 'PoP! Goes My Heart'. Even though you are literally on the other side of the world, I still feel your friendship. Thank you for being there.

To all my colleagues on the 9th floor, thank you for providing me with such a supportive, intellectual, and "gezellige" work environment. You have created a special place. Your work ethic is admirable, as is your play ethic. I have always felt welcomed and respected on the 9th. To my delight, you have always welcomed me back with a genuine smile, even – or perhaps especially – when I had been gone for a long time. Admittedly, it is of course possible that I may have exaggerated these smiles in my mind, given my strong priors about your likability. Genuine smiles or not, you always took an interest in how I was doing and I thank you for that. My days on the 9th floor would not have been the same without you. I am thankful to all of you and will mention some of you by name.

Ap, thank you for emphasizing the importance of holidays and fun and for being willing to offer me a PhD position if I had not been awarded the NWO grant. Johan, thank you for that same willingness and for enthusiastically storming our office from time to time. Also, please keep writing your amazing Sinterklaas poems. They always put a smile on my face. Harm, thank you for your dry-wit humor and for promoting healthy eating habits. Gijs, thank you for your kindness and warmth and for keeping person perception research alive on the 9th (now 4th) floor. Thijs, thank you for always greeting me enthusiastically! Cor, thank you for your hilarious photo contest presentations and for your invaluable work on the teaching front. Maerten, thank you for gently pushing me to keep teaching. Hein, thank you for your challenging questions in lab meetings. Martijn, Barbara, Thijs, and Mattheis, thank you for rocking those

cowboy boots, toe socks, leopard pants, and that flamingo suit. You make the 9th floor a more colorful place. Roel and Luuk, thank you for your many fun distractions every time you walked into our office. Sterre and Thomas, thanks for the many conversations and delicious summer BBQs.

To my fellow (former) PhD students. André, Carmel, Gijs, Kai Qin, Lieke C., Maikel, Maitta, Reine, Rik, Sanne, Tom, and Thijs, thank you for your warm welcome in PhD life during the early days of my PhD project. Asteria, Bernice, Diamantis, Gesa, Iris, Jeroen, Julian, Kim Lien, Lieke S., Mike Z., Niklas, Peiying, Piotr, Sari, Shuang, Ted, Tiziano, Tjits, Xiaojing, Yuxi, and Zhang, thank you for making PhD life more wonderful. I miss our PhD dinners, campus walks, drinks, office and party conversations, mental support, and most of all your lovely smiles. Maitta and Asteria, thank you for your serene presence and travel advise for Barcelona and Bali. Julian and Zhang, thank you for sharing your optimism and research related wisdom. Kim Lien, Lieke, Peiying, Sari, Shuang, and Xiaojing, thank you for your heartfelt sympathy and for making office life that much funnier. I love chatting with you. Yuxi, in my mind I keep seeing you entering the elevator with your sports bag and a beaming smile on your face. Thank you for spreading happiness. I hope life treats all of you well.

Marjo, Madelon, Monique, Meta, Ronny, en Rob Gommans, dank voor jullie waardevolle ondersteuning. Jullie maken het leven van BSI onderzoekers makkelijker.

Thanks to all members of the Person Perception group, the Behavior Regulation group, the Social Mind group, and the Utrecht RC group for inspiring lab meetings and your valuable input.

Thanks to my awesome colleagues of the PhD Platform, Carla, Nessa, Adam, Jeroen, Lian, Erik, Ciska, Loes, Elke, Evelien, Bert and Eliane. Together we managed to make meaningful contributions to the careers of BSI PhD students while having loads of fun as well. I have fond memories of the homemade cocktails.

To all colleagues outside the 9th floor and outside Nijmegen: thank you for the inspiring conferences and workshops, amazing summer school experiences, interesting conversations, good times, and for your kindness.

Thanks to the people behind Prolific and Gorilla. Your amazing online services helped to make the research in this dissertation possible.

To those who I have not mentioned by name, but who supported me nevertheless: thank you!

Mijn dank gaat ook uit naar alle zorgverleners die zich hebben ingezet (en nog steeds inzetten) om me te helpen herstellen. Drie stapjes naar voren, twee stapjes terug. Uiteindelijk ga je toch vooruit.

Aan mijn lieve vrienden Babs, Joni, Lianne, Sarah, Eli Elise, Femke, Richelle, Mandy, Anita, Saskia, Kim, Giel, en de 'Kraaytjes'. Dank jullie wel voor alle lieve steun en vermakelijke afleidingen tijdens mijn PhD project. Dansen onder de maan, gezellige logeerpartijtjes, wandelen in de natuur, diepgaande gesprekken, dansjes in de kroeg, high tea middagen, yoga sessies, SinterKerst cadeauspelletjes, watergevechten, en lieve kaartjes, appjes, en belletjes. Jullie verrijken mijn leven. Dank voor jullie vriendschap.

Sam en Floor, mijn reisgezellen in het leven. Soms van heel dichtbij en soms vanuit de verte hebben jullie alle ontwikkelingen meegekregen in mijn leven. Nu heeft jullie kleine zusje een boekje geschreven. Dank jullie wel voor alle steun. Van 4x helpen verhuizen tot het bieden van een luisterend oor en van samen spelletjes spelen tot shoppen voor een trouwjurk. Floor, ik hoop dat we nog jarenlang op weekendjes weg blijven gaan. Natuurlijk ook mijn dank voor alle lieve steun van Nienke en Erik, en van Mick, Roos, en Fen, de leukste kleine afleiders.

Martin en Saskia, ook jullie bedankt voor alle steun de afgelopen jaren. Alle bezoekjes, mailtjes, belletjes, brieven, goedgevulde heen-en-weer tassen, en prachtige schilderwerkjes. Jullie interesse en vertrouwen in mij verwarmt me.

Harry en Petra, mijn veilige haven. Lieve pap en mam, wat ben ik gelukkig om jullie nog steeds bij me te hebben. Ik vind jullie prachtige mensen en voel zoveel liefde en dankbaarheid voor jullie. Jullie staan altijd voor me klaar, of ik nu alleen even wat gezelligheid nodig heb of zelfs een plek om elk weekend te kunnen herstellen. Bedankt voor al jullie steun en lieve zorgen tijdens het ziek zijn, en ook tijdens het niet ziek zijn. Misschien heb ik geluk omdat jullie al op twee kinderen hadden kunnen oefenen, maar ik sta nog steeds versteld van hoe fantastische ouders jullie zijn. Ik hoop dat jullie nog decennia vrolijk blijven rondfietsen en dat ik nog lange tijd van jullie kan blijven leren en genieten. Bedankt voor alles.

Béla, mijn grote liefde, rots in de branding, en paranimf. Jij bent als een schat die ik gevonden heb en die wonderbaarlijk genoeg nog bij me wil blijven ook. Wat een bofkont ben ik om zo'n reisgenoot te hebben gevonden als jij. Mijn eerste "boek" schreef ik voor jou ('Liefde is… 2 poebeliewoepsies') en mijn tweede is nu afgerond mede dankzij jou. Dat ik dit proefschrift heb kunnen schrijven heb ik voor een groot deel te danken aan jouw gouden steun, geduld, zorgzaamheid, magische ontspanningsknuffels, intelligente vragen, luisterend oor, strenge toespraakjes, vertrouwen, diepe gesprekken, toejuichingen, liefde, kookkunsten, en humor. Jij moest dan ook wel mijn paranimf worden! Ik vind het knap hoe je mij tegelijkertijd steunt en de vrijheid geeft. Je bent precies wat ik nodig heb. Dankjewel voor jou, lieve schat. Ik kijk uit naar ons volgende avontuur waarin we samen verder reizen als man en vrouw.

Curriculum Vitae

Lin Fiene Jansen was born on the 4th of December 1988 in Arnhem, the Netherlands. She graduated cum laude from the bilingual pre-university secondary education program of Lorentz Lyceum in Arnhem. In 2008, she started studying Psychology at Radboud University in Nijmegen. During her bachelor, she completed the Honors Program of Psychology that allowed her to visit Princeton University (USA) as part of a research internship. In 2011, she obtained her Bachelor of Science degree (cum laude) in Psychology and was admitted to the Research Master Behavioural Science at Radboud University. During her master, she monitored the quality of the curriculum as a member of the program committee and enjoyed working as a student research assistant. Furthermore, she was admitted to the Honors Program Beyond the Frontiers that allowed her to complete a research internship at the University of Glasgow (UK). In 2013, Lin obtained her Master of Science degree (cum laude) and was awarded the NWO Research Talent grant that allowed her to start working as a PhD candidate at the Behavioural Science Institute (BSI) of Radboud University. During her PhD project, she took on several roles, such as mentor to research master students, teacher in the Psychology bachelor, BSI lab schedule manager, and PhD representative in the BSI PhD platform. She also attended the EASP summer school of 2014 in Lisbon (Portugal). Over the years, she was forced to pause her PhD project twice as she was coping with a long-term illness. She completed her PhD project on the circumstances under which people (do not) read information into faces under the supervision of prof. dr. Daniël Wigboldus, prof. dr. Rob Holland, and dr. Ron Dotsch, which resulted in the present dissertation. Currently, Lin works as a teacher in the Psychology bachelor at Radboud University and is training to further improve her coaching and training skills.

Behavioural
Science
Institute

Radboud University