Testing perceivers' accuracy and accuracy awareness when forming personality impressions from faces

Jaeger, Bastian; Sleegers, Willem; Stern, Julia; Penke, Lars; Jones, Alex L.

DOI:
10.31234/osf.io/np9ec

Publication date:
2023

Document Version
Early version, also known as pre-print

Testing Perceivers' Accuracy and Accuracy Awareness When Forming Personality Impressions From Faces

Bastian Jaeger[1], Willem W. A. Sleegers[1], Julia Stern[2], Lars Penke[3], Alex L. Jones[4]

[1]Tilburg University, [2]University of Bremen, [3]University of Göttingen, [4]Swansea University

Draft version: 9 November 2023

This paper is currently undergoing peer review. Comments are welcome.

Word count: 12,131

Author Note

Bastian Jaeger and Willem W. A. Sleegers, Department of Social Psychology, Tilburg University, The Netherlands; Julia Stern, Department of Psychology, University of Bremen, Germany, Lars Penke, Department of Psychology, University of Goettingen, Germany; Alex L. Jones, School of Psychology, Swansea University.

Correspondence concerning this article should be addressed to Bastian Jaeger, Department of Social Psychology, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: b.jaeger@uvt.nl.

Abstract

People spontaneously judge others' personality based on their facial appearance and these impressions guide many important decisions. Although the consequences of personality impressions are well documented, studies on the accuracy of personality impressions have yielded mixed results. Moreover, relatively little is known about people's accuracy awareness (i.e., whether they are aware of their judgment accuracy). Even if accuracy is generally low, awareness of accuracy would allow people to rely on their impressions in the right situations. In two studies (one preregistered), we estimated perceivers' accuracy and accuracy awareness when forming personality impressions based on facial photographs. Our studies have three crucial advantages as compared to previous studies (a) by incentivizing accuracy and accuracy awareness, (b) by relying on substantially larger samples of raters ($n_{Study\ 1}$ = 223, $n_{Study\ 2}$ = 423) and targets ($k_{Study\ 1}$ = 140, $k_{Study\ 2}$ = 1,260 unique pairs with 280 unique targets), and (c) by conducting Bayesian analyses to also quantify evidence for the null hypothesis. Our findings suggest that face-based personality impressions are not accurate, that perceivers lack insight into their (in)accuracy, and that most people overestimate their accuracy.

Keywords: personality impressions; face perception; accuracy; accuracy awareness; confidence

Testing Perceivers' Accuracy and Accuracy Awareness When Forming Personality Impressions From Faces

People form impressions of others' personality based on their facial appearance alone (Todorov, Olivola, et al., 2015). These impressions are formed within a few hundred milliseconds (Willis & Todorov, 2006) and can be very consequential, as people rely on them, next to other cues, to make important decisions (Olivola et al., 2014). Facial impressions have been shown to influence voting decisions (Olivola & Todorov, 2010a), criminal sentencing (Berry & Zebrowitz-McArthur, 1988; Porter et al., 2010), personnel selection (Gomulya et al., 2017; Li et al., 2017), consumer behavior (Duarte et al., 2012; Jaeger, Sleegers, et al., 2019), and many other choices. Although it is difficult to determine exactly *how* large the influence of facial impressions is compared to the many other cues that people rely on in decision-making, several studies suggest effects of facial impressions are often not trivial. For example, Graham and colleagues (2016) found that a one standard deviation increase in perceived competence was associated with an eleven to fourteen percent increase in the starting salary of CEOs (after controlling for various other factors). Effects of facial impressions are also relatively persistent. People rely on facial impressions even when they have access to more diagnostic information (Jaeger, Evans, et al., 2019; Rezlescu et al., 2012) or when they are instructed to ignore facial appearance (Chua & Freeman, 2021; Hassin & Trope, 2000; Jaeger et al., 2020; Shen & Ferguson, 2021).

The notion that people rely on personality impression from faces is well-supported by the extant literature, but the question of how accurate these impressions are and, relatedly, whether reliance on them should be considered problematic, remains debated (Bonnefon et al., 2015; Todorov, Funk, et al., 2015). Some highlight evidence for above-chance accuracy when judging various traits based on a person's facial appearance (Alper et al., 2020; De Neys et al., 2017; Lin et al., 2018; Penton-Voak et al., 2006). This would imply that a person's facial appearance is a valid source of information and reliance on personality impressions is not necessarily problematic, especially when more valid cues are not available. Others point to null findings or argue that accuracy is so low that reliance on face-based personality impressions should be considered as a source of bias in decision-making (Olivae et al., 2014; Todorov, Funk, et al., 2015; Vogt et al., 2013). Here, we argue that it is not only important to understand how accurate people are in inferring personality traits from faces, but also whether they are aware of their

accuracy (i.e., whether they know when their judgments are more or less accurate). Even if people's impressions are inaccurate most of the time, reliance on them could still be justified if people can discriminate between instances in which their impressions are more accurate and can be relied upon, and instances in which their judgments are inaccurate and should not be relied upon. That is, accuracy awareness can foster adaptive reliance on personality impressions, even if accuracy is relatively low.

Accuracy of Personality Impressions from Faces: Theoretical Accounts

The question of whether people's personality judgments from faces correspond to targets' personality has received considerable attention in the literature (Borkenau & Liebler, 1992; Naumann et al., 2009; Penton-Voak et al., 2006). Several authors have noted that the ability to infer others' traits based on their facial appearance, which is a readily available cue in everyday social interactions, would be very beneficial (DeSteno et al., 2012; Verplaetse et al., 2007). This view fits with functionalist approaches to social perception, which highlight that social-perceptual processes likely evolved to help people navigate their social environment (Adams et al., 2017; Zebrowitz, 2012). Therefore, the tendency to spontaneously form and rely on personality impressions could have evolved because an individual's facial appearance is a somewhat valid indicator of their personality.

Three main counter-arguments to this line of reasoning have been made. First, although people show some consensus when making face-based personality judgments, there are substantial individual differences (Hehman et al., 2019). When perceivers show complete agreement in their impressions of different targets, 100% of the variance in their judgments is explained by the specific target they are judging. However, recent studies have shown that only 10-25% of the variance in trait judgments is explained by the appearance of the person being judged (Hehman et al., 2017). Most of the variance is explained by who is making the judgment (i.e., characteristics of the perceiver) and by the interaction between target and perceiver. This observation speaks against the argument that personality impressions arise from a widely shared evolved cognitive mechanism. More importantly, these results suggest that the accuracy of personality impressions can only be very limited at best, as personality judgments of various perceivers can only be accurate if they are reliable across different perceivers.

Second, the exact mechanism that would enable accurate personality impressions has remained somewhat unclear. A prerequisite for accurate personality impressions from faces is

the presence of valid cues (i.e., facial features that correlate with personality; Funder, 2012). Some have proposed that facial width-to-height ratio is a valid indicator of various traits, including trustworthiness (Stirrat & Perrett, 2010), aggressiveness (Carré et al., 2009), and sociosexuality (Bird et al., 2017) because both facial appearance and personality are determined by common biological factors (e.g., testosterone). However, many reported links between facial width-to-height ratio and various traits have failed to replicate in larger samples (Burris & Edwards, 2017; Deaner et al., 2012; Kordsmeyer et al., 2019; Kosinski, 2017; Wang et al., 2019).[1] Studies on the validity of other facial cues, such as attractiveness (Feingold, 1992; Langlois et al., 2000; Little & Perrett, 2007; Nestler et al., 2012), as indicators of personality traits have also yielded mixed results. Others have proposed that people who are treated differently because of their facial appearance may end up developing the exact trait that was incorrectly ascribed to them, leading to a self-fulfilling prophecy (Slepian & Ames, 2015). For example, people who are perceived as unapproachable and therefore avoided due to their facial appearance might actually become more antisocial. This account predicts that people with a similar facial appearance will be treated in similar ways and should therefore show similar personalities. However, studies comparing personality traits of genetically unrelated individuals with very high levels of facial similarity found little support for this prediction (Segal, 2013; Segal et al., 2013, 2018). Thus, clear evidence on which facial cues would allow perceivers to make accurate personality judgments has failed to emerge thus far.

Third, overgeneralization theory offers a plausible alternative account of people's tendency to judge others' personality based on their facial appearance even if these judgments are not accurate (Oosterhof & Todorov, 2008; Todorov, Olivola, et al., 2015; Zebrowitz, 2012, 2017; Zebrowitz et al., 2003). This account holds that personality impressions from faces are by-products of adaptive social-cognitive systems that extract relevant information from faces. For example, facial expressions, such as smiles, communicate important social information (Van Kleef, 2010), which leads perceivers to be especially attuned for facial cues that indicate the presence of a smile. This sensitivity causes people to perceive smiles and other emotional

[1] Relatedly, recent work also challenges the idea that people rely on facial width-to-height ratio when forming impressions (Durkee & Ayers, 2021). Many facial cues, including facial width-to-height ratio, are intercorrelated, making it difficult to isolate the unique effect of a certain cue on impression formation. Studies that examined a wider range of facial features did not find unique associations between facial width-to-height ratio and impressions. (Jaeger & Jones, 2021; Windmann et al., 2021).

expressions in faces that merely *resemble* the emotional expression (Said et al., 2009). As a consequence, judgments of people who display a smile (e.g., that they are warm and extraverted) are overgeneralized to people whose natural facial appearance resembles a smile. Studies examining the determinants of personality impressions have yielded support for these predictions, showing that perceivers rely on resemblances to emotional expressions when judging others (Adams et al., 2012; Jaeger & Jones, 2021; Windmann et al., 2021).

Accuracy of Personality Impressions from Faces: Empirical Evidence

While the plausibility of accurate personality impressions based on targets' facial appearance is still debated, empirical evidence on the topic has started to accumulate. In a typical study, a sample of participants (i.e., the targets) is photographed and asked to complete a personality measure. Targets are usually instructed to maintain a neutral facial expression and are photographed against a uniform background (Naumann et al., 2009; Nestler et al., 2012; Penton-Voak et al., 2006). The photographs are then shown to a second sample of participants (i.e., the perceivers) who rate the targets' personality. To test if perceivers form accurate impressions of targets' personality, correlations between their ratings and target's self-reported scores, often referred to as trait accuracy, are examined. This shows to what extent perceivers are able to discriminate between different targets on a given trait (Biesanz, 2010)[2]. Although most investigations relied on targets' self-reported personality as the accuracy criterion, some also solicited informant ratings and averaged self-reported and informant ratings into a composite accuracy criterion (Ames et al., 2010; Naumann et al., 2009). Studies also varied in how personality judgments were assessed. In some studies, perceivers evaluated targets with the same personality questionnaire that was also used by targets (Naumann et al., 2009; Nestler et al., 2012). In others, perceivers' judgments were assessed with simpler rating scales, such as a single

[2] Research on the accuracy of personality judgments sometimes distinguishes between different types of accuracy (Back & Nestler, 2016; Biesanz, 2010; Hall et al., 2018; Letzring, 2008; Letzring et al., 2021). Accuracy can be examined at the inter-target or intra-target level. *Trait accuracy* refers to perceivers' ability to accurately judge the relative levels of different targets on a given trait (e.g., Charlie correctly judges that Alfred is more extraverted than Russell), whereas *profile accuracy* refers to perceivers' ability to accurately judge the relative levels of different traits in a given target (e.g., Charlie correctly judges that Alfred is more extraverted than agreeable). *Normative accuracy* refers to the association between perceivers judgments and the average characteristics of targets (e.g., the average level of extraversion across targets), whereas *distinctive accuracy* refers to the association between perceivers' judgments and targets' unique characteristics (e.g., the extent to which they differ from the average person on extraversion). Following the majority of previous investigation on the accuracy of personality impressions from faces (Ames et al., 2010; Borkenau et al., 2009; Naumann et al., 2009; Satchell et al., 2019), we focus on trait accuracy, which is also closest to lay definition of accuracy. We examine perceivers' trait accuracy for each of the Big Five dimensions and their mean trait accuracy across the five dimensions.

item accompanied by a description (Alper et al., 2020) or a single bipolar item (Borkenau et al., 2009).

Differences in methodological and statistical approaches make it difficult to integrate findings of previous studies. Perhaps the only strong conclusion that can be drawn based on existing studies is that the accuracy of facial impressions is unclear due to the many inconsistent findings in the literature. Extraversion judgments usually show the highest levels of accuracy in personality rating tasks (Kenny & West, 2008; Letzring et al., 2021). Some studies also found significant levels of accuracy for extraversion impressions based on face images. For example, Naumann and colleagues (2009) found a correlation of $r = .29$ between perceivers' extraversion impressions and targets' extraversion scores. Similar results were obtained in other studies (Nestler & Back, 2013; Penton-Voak et al., 2006). However, others found no evidence for accuracy (Ames et al., 2010; Shevlin et al., 2003).

Results are even more inconsistent for the other Big Five dimensions. For example, Nestler and colleagues (2012) found evidence for accurate impressions of emotional stability, but only in one of their studies in which photos of female targets were used. These findings conflict with the results of Penton-Voak and colleagues (2006), who analyzed male and female separately and found significant levels of accuracy for male but not female targets. Ames and colleagues (2010) found evidence for accuracy in a sample with both male and female targets, but two other studies did not (Naumann et al., 2009; Shevlin et al., 2003). Similar inconsistent findings have emerged for impressions of openness, conscientiousness, and agreeableness.

Face prototypes are a popular alternative method to assess the accuracy of facial impressions (Penton-Voak et al., 2006). In this approach, photos of individuals that score particularly high or low on a certain dimension (e.g. extraversion) are selected and morphed to create face prototypes (e.g., an extraverted and an introverted face prototype). The two prototypes are presented next to each other and perceivers judge which of the two scores higher on the dimension of interest. If perceivers judge the high-extraversion prototype as more extraverted than the low-extraversion protype at a rate higher than expected by chance (i.e., 50%), then this is taken as evidence for the accuracy of extraversion impressions. Studies using this approach have also yielded mixed results. Although some found evidence for accurate extraversion impressions (Kramer & Ward, 2010; Little & Perrett, 2007; Penton-Voak et al., 2006), others did not (A. L. Jones et al., 2012) or the evidence was inconsistent across studies

(Alper et al., 2020). Findings are similarly inconsistent for impressions of conscientiousness and emotional stability, but more consistent for impressions of openness and agreeableness. None of the relevant studies we were able to identify found evidence for accurate impressions of openness, whereas all found evidence for accurate impressions of agreeableness (Alper et al., 2020; A. L. Jones et al., 2012; Kramer & Ward, 2010; Little & Perrett, 2007; Penton-Voak et al., 2006).

It should be noted that the validity of the prototype method has been criticized on various grounds (Bovet et al., 2022; DeBruine, 2020; A. L. Jones & Jaeger, 2019). Prototypes are often based on a few targets (Little & Perrett, 2007; Penton-Voak et al., 2006), making it questionable whether they are reliable representations of the average facial appearance associated with a given trait (Bovet et al., 2022). Even if the procedure succeeds in distilling the average facial appearance associated with a given trait, the resulting prototypes are not externally valid stimuli. By presenting the two prototypes next to each other, even small differences in facial appearance, which perceivers may not detect in everyday life, are highlighted which can artificially inflate judgment accuracy. Overall, it is questionable how much personality judgments of prototypes can tell us about everyday personality judgments.

In some studies, perceivers judged targets' personality based on somewhat richer static stimuli, such as selfies taken from social media (Qiu et al., 2015). Evidence from these studies is less relevant for understanding the accuracy of face-based personality impressions because the stimuli also contained other cues that perceivers could rely on to form impressions. Although richer stimuli generally lead to more accurate judgments (Funder, 2012; Krzyzaniak et al., 2019), evidence for the accuracy of personality impressions based on stimuli that reveal various aspects of a person's appearance is again somewhat mixed. Studies in which targets were allowed to adopt any facial expression while being photographed (Borkenau et al., 2009; Naumann et al., 2009) and studies that filmed (Borkenau & Liebler, 1992) or photographed targets sitting at a table from the waist up (Beer, 2014; Beer & Watson, 2010) found evidence for accurate impressions of extraversion. Extraversion impressions based on social media profile photos (Stopfer et al., 2014) and selfies that were submitted by targets were also somewhat accurate (Satchell et al., 2019). However, another study that examined impressions based on profile photos downloaded from social media found no evidence for accuracy (Qiu et al., 2015). Results are even less consistent for the other Big Five dimensions (Beer, 2014; Beer & Watson, 2010;

Borkenau et al., 2009; Borkenau & Liebler, 1992; Naumann et al., 2009; Qiu et al., 2015; Satchell et al., 2019; Stopfer et al., 2014). Although each study found significant levels of accuracy for at least one dimension, evidence on *which* dimensions can be judged accurately varies substantially from study to study (Beer, 2014; Beer & Watson, 2010).

Overall, the state of the evidence on the accuracy of personality impressions from faces is mixed with many inconsistent findings. The majority of published studies found above-chance accuracy for extraversion impressions (Naumann et al., 2009; Penton-Voak et al., 2006), which is in line with previous work on personality judgments in more information-rich environments, such as brief face-to-face interactions (Kenny & West, 2008; Letzring et al., 2021). However, some studies did not find evidence for accuracy in spite of similar methods and sample sizes (Ames et al., 2010; Shevlin et al., 2003). Results are even more inconsistent for impressions of openness, conscientiousness, agreeableness, and emotional stability. These inconsistencies not only emerged in studies where perceivers judged targets based on facial photographs, but also in studies where perceivers judged face prototypes and somewhat richer stimuli such as selfies.

Accuracy Awareness

Although a substantial body of prior work has focused on elucidating the accuracy of face-based personality impressions, few studies have examined accuracy awareness. Accuracy awareness is usually measured by examining the relation between how accurate perceivers' impressions are (e.g., whether their personality impressions of targets correspond to targets' self-reported personality scores) and how accurate perceivers *think* their impression are (Ames et al., 2010). Accuracy awareness can be assessed at different levels. Targets vary in their trait expressiveness (Funder, 2012; Human et al., 2021), which means that the same perceiver may show different levels of accuracy across different targets. Perceivers may also be aware of this fluctuation in the accuracy of their judgments across different situations. We refer to this as *within-perceiver accuracy awareness*. It is also plausible that perceivers' accuracy awareness is not so fine-grained. Although perceivers may not know if their judgments are more or less accurate when judging a specific target, they may know whether their judgments are generally more or less accurate. That is, due to differences in ability (de Vries et al., 2021) or motivation (Biesanz & Human, 2010; Capozzi et al., 2020) some perceivers could be better judges of personality than others and perceivers may be aware of this. We refer to this as *between-perceiver accuracy awareness*.

Direct evidence on perceivers' accuracy awareness when forming face-based personality impressions is sparse. Although Borkenau and colleagues (2009) did not analyze between-perceiver or within-perceiver accuracy awareness directly, they found that perceivers were most confident when judging extraversion, which was also the only trait for which judgments were significantly related to self-reported scores. Ames and colleagues (2010) conducted a more comprehensive analysis of accuracy awareness. Perceivers judged the personality of 21 targets and indicated their confidence in each judgment. There was no significant correlation between perceivers' judgment accuracy and their confidence across the 21 targets. Moreover, perceivers were not significantly more confident when judging targets that were also perceived more accurately. In short, this study did not find evidence for either between-perceiver or within-perceiver accuracy awareness.[3]

Additional support for a lack of between-perceiver accuracy awareness is provided by studies that only examined confidence in the accuracy of personality impression from faces. Research on lay beliefs in physiognomy (i.e., beliefs in the manifestation of personality traits in facial appearance) suggest that people are relatively confident in their ability to judge others' personality based on their looks (Jaeger, Evans, et al., 2022; Realo et al., 2003; Suzuki et al., 2017). A survey among Japanese and U. S. American participants showed that many people believe that they can infer various personality traits from a person's face. For example, 47% of Japanese respondents and 69% of American respondents indicated that they can tell how kind a person is from looking at their face. These results could be explained by the fact that many people expect their judgments to be accurate when they can rely not only on a person's stable facial features, but also on dynamic facial characteristics such as smiles and other facial expressions. Yet, even when participants viewed photographs of faces with a neutral expression, confidence in the accuracy of personality judgments was relatively high (Hassin & Trope, 2000; Jaeger, Evans, et al., 2022). Around 50% of Dutch students and U. S. American MTurk workers indicated that they can learn something about the personality of a stranger just from looking at

[3] Some evidence for accuracy awareness was found when perceivers judged targets based on more than their facial appearance. Ames and colleagues (2010) showed perceivers 60-second videos of targets who participated in a mock interview. They did not find evidence for between-perceiver accuracy awareness, but significant evidence for within-perceiver awareness in one of the two studies (the effect was only marginally significant in one study). Specifically, perceivers with moderate confidence were more accurate than perceivers with low confidence. A similar pattern was found in another study in which participants engaged in a 3-minute conversation before rating each other (Biesanz et al., 2011).

their passport photo (i.e., a face with a neutral expression). These high levels of confidence in face-based personality impressions are at odds with research showing limited evidence for accuracy, suggesting that people may be overconfident when forming personality impressions. This can only be taken as indirect support for a lack of accuracy awareness, as these studies did not measure the accuracy of perceivers' impressions. Overall, despite its theoretical importance, few studies have examined accuracy awareness in personality impressions.

The Current Studies

Here we present the results of two studies (one preregistered) that tested perceivers' accuracy and accuracy awareness when forming personality impressions from faces. We compare whether perceivers' impressions based on facial photographs are related to targets' self-reported personality traits. Perceivers also indicate how confident they are in the accuracy of their impressions, and we test whether their confidence is related to their actual accuracy. We examine both within-perceiver and between-perceiver accuracy awareness.

Our studies improve on previous work on this topic in three critical ways. First, incentives have been shown to improve accuracy and accuracy awareness in various judgments tasks (e.g., Botvinick & Braver, 2015; Lebreton et al., 2018). Prior studies examining personality judgments based on richer social stimuli (e.g., video interviews) found that accuracy rates increased when perceivers spent more time looking at the target (Capozzi et al., 2020) and when perceivers were instructed to make judgments as accurately as possible (Biesanz & Human, 2010), which suggests that motivation may increase personality judgment accuracy. It is possible that the absence of judgment accuracy observed in many previous studies was due to perceivers' low motivation and that perceivers' would show higher levels of accuracy if their judgments are tied to some meaningful outcome (as they also tend to be in everyday life). Yet, no prior studies on face-based personality impressions financially incentivized perceivers' judgments or their judgment confidence. We therefore designed an incentive-compatible judgment task in which perceivers are incentivized to provide accurate personality judgments and accurate estimates of their judgment accuracy.

Second, many previous findings on the accuracy of personality impressions from faces are based on relatively small samples with fewer than 50 raters (e.g., Borkenau et al., 2009; Naumann et al., 2009; Stopfer et al., 2014), fewer than 50 targets (e.g., Kramer & Ward, 2010; Nestler et al., 2012; Shevlin, 2003), or both (e.g., Ames et al., 2010; Little & Perret, 2007).

Crucially, the only other study that examined both accuracy and accuracy awareness relied on a sample of 25 perceivers and 21 targets (Ames et al., 2010). It is difficult to estimate to what extent prior investigations were underpowered to detect meaningful effect sizes given the more complex study designs and a lack of consensus on what would constitute the smallest effect size of interest. However, large samples of raters and targets are also crucial for precision in estimating accuracy and accuracy awareness and for testing whether results generalize beyond a specific set of raters and stimuli (Judd et al., 2012). We therefore rely on much larger samples of raters ($n_{Study\ 1}$ = 223, $n_{Study\ 2}$ = 423) and targets ($k_{Study\ 1}$ = 140, $k_{Study\ 2}$ = 1,260), analyzing more than 60,000 personality judgments in total.

Third, in light of the limited and inconsistent evidence in favor of accuracy and accuracy awareness, it is plausible that perceivers show neither when forming personality impressions from faces. Yet, existing studies have exclusively focused on statistical methods that cannot provide evidence for such a null hypothesis. In other words, it is unclear whether previous studies did not find accuracy levels that significantly differed from chance because the accuracy of personality impressions is in fact at chance level, which is plausible given low levels of consensus (Hehman et al., 2017) and the thus far unsuccessful search for valid facial cues reflecting an individual's personality (Kosinski, 2017; D. Wang et al., 2019), or because studies lacked the statistical power to detect modest levels of accuracy, which is also plausible given the small samples of raters and targets. We therefore report the results of Bayesian analyses (alongside frequentist statistics) that quantify evidence in favor of the null hypothesis (Wagenmakers, 2007). This allows us to estimate the extent to which our results are in line with chance-level accuracy and accuracy awareness.

Study 1

In Study 1, we measured perceivers' accuracy and accuracy awareness when judging how targets score on the Big Five personality traits. We examined trait accuracy scores for each dimensions and mean trait accuracy across the five dimensions. Perceivers saw facial photographs of female targets displaying a neutral facial expression and indicated (a) their personality impressions and (b) their confidence in the accuracy of their impressions. We examined whether perceivers' ratings were associated with targets' self-report scores, and whether perceivers were more confident in their ratings when their ratings were actually more accurate. Both accuracy and accuracy awareness were incentivized independently.

Methods

Participants. We recruited 232 first-year psychology students from a Dutch university who completed the study in return for partial course credit and two chances to win a €50 voucher. It took participants approximately 8 minutes to complete the study. The sample size was determined by how many participants completed the study within two weeks. Data from 4 participants (1.72%) who indicated that the stimuli did not load properly and from 5 participants (2.19%) who always provided the same response across all trials were excluded, leaving a final sample of 223 participants (M_{age} = 20.3 years, SD_{age} = 2.3; 67.71% female, 31.39% male, 0.90% other).[4]

Stimuli. We used facial photographs of 141 female students from a German University (18-34 years old). Photographs were taken with a digital camera (Canon EOS 350D). Targets stood in front of a white background and were instructed to display a neutral facial expression. Standing position, lighting, and distance were standardized (for more details, see Jünger et al., 2018). Personality was assessed with the 44-item Big Five Inventory (O. P. John et al., 2008). Targets indicated their agreement with each statement on a five-point scale. Average scores on the five dimensions showed acceptable to good internal consistency (openness: α = .82, conscientiousness: α = .80, extraversion: α = .84, agreeableness: α = .74, emotional stability: α = .77). We created 7 image sets, each containing 20 face images. One random image was dropped in order to create an even number of stimuli (k = 140).

Procedure. Perceivers were randomly assigned to one image set. We measured personality impressions by asking perceivers to judge the person in the photo on each of the Big Five dimensions. In line with previous work (Borkenau et al., 2009, Little & Perrett, 2007; Satchell et al., 2019, see also Alper et al., 2020), perceivers rated targets on one dimension at a time using a single item that ranged from 1 (*not* [trait] *at all*) to 5 (*extremely* [trait]). Thus, both self-reported and judged personality were assessed with five-point scales. At the beginning of the study, each dimension was described using two trait adjectives from the Ten-Item Personality Inventory (Gosling et al., 2003). For example, for conscientiousness, participants read: "A person who scores low on conscientiousness is disorganized and careless. A person who scores high on conscientiousness is dependable and self-disciplined." These descriptions were also shown during each trial. Each time participants were asked to rate a target on a specific

[4] We obtained similar results when including these data in our analyses.

dimension, they saw the description of the dimension. Perceivers rated 20 targets on 5 different dimensions for a total of 100 trials. After each trait rating, perceivers also indicated their confidence. Similar to previous work on facial impressions (Dotsch & Todorov, 2012; Mattarozzi et al., 2015), perceiver indicated their confidence in the accuracy of their ratings on a scale that ranged from 1 (*not confident at all*) to 9 (*extremely confident*). Personality and confidence ratings were not time-constrained. The order in which the faces and personality dimensions were displayed was randomized. On average, each face was judged by 17-39 unique raters (M = 31.86, SD = 4.27).

We computed intraclass correlation coefficients (Shrout & Fleiss, 1979) to estimate how much variance in personality impressions was due to target effects (i.e., consensus) and perceiver effects (i.e., assimilation; Kenny, 1994). Across stimulus sets, consensus estimates ranged from ICC(2,1) = .135 to ICC(2,1) = .284 for openness judgments, from ICC(2,1) = .030 to ICC(2,1) = .217 for conscientiousness judgments, from ICC(2,1) = .160 to ICC(2,1) = .248 for extraversion judgments, from ICC(2,1) = .103 to ICC(2,1) = .222 for agreeableness judgments, and from ICC(2,1) = .084 to ICC(2,1) = .267 for emotional stability judgments (all ps < .001). Assimilation estimates ranged from ICC(2,1) = .070 to ICC(2,1) = .177 for openness judgments, from ICC(2,1) = .102 to ICC(2,1) = .171 for conscientiousness judgments, from ICC(2,1) = .076 to ICC(2,1) = .146 for extraversion judgments, from ICC(2,1) = .081 to ICC(2,1) = .166 for agreeableness judgments, and from ICC(2,1) = .077 to ICC(2,1) = .139 for emotional stability judgments (all ps < .001).

Both accuracy and accuracy awareness were incentivized. Perceivers were informed that the person with the most accurate ratings (i.e., the person with the strongest correlation between personality ratings and the accuracy criterion) and the person with the highest accuracy awareness (i.e., with the strongest correlation between accuracy and confidence) would each be rewarded with a €50 voucher for an online retailer.

Analysis strategy. All analyses were conducted in R (R Core Team, 2021). Multilevel regression models were estimated with the *lme4* package (Bates et al., 2015) and p-values were computed with the *lmerTest* package (Kuznetsova et al., 2016). As observations were nested within perceivers and within targets, all models included random intercepts for perceivers and targets. We also included random slopes per perceiver and target to model variation in trait accuracy and accuracy awareness (details on how accuracy and accuracy awareness were tested

are reported in the Results section). When testing for mean trait accuracy and accuracy awareness across the Big Five dimensions (rather than for each dimension separately), our models also included a random intercept per dimension.

We report the results of Bayesian analyses alongside frequentist statistics. We computed Bayes factors for correlation coefficients and *t*-tests using the *BayesFactor* package with default priors (Morey & Rouder, 2018). We also explored the robustness of our results by implementing different priors (see Supplemental Materials). To compute Bayes factors for coefficients in multilevel regression models, we followed the approach outlined by Wagenmakers (2007). We estimated models with and without the fixed effect of interest and computed the Bayesian information criterion (BIC), an indicator of model fit, for both models. We compared the BICs of both models to quantify the extent to which the fixed effect of interest improved model fit. Following Wagenmakers (2007), we converted this measure to an approximation of the Bayes factor with the following formula: $BF_{10} \approx exp\left(\frac{BIC(H_0) - BIC(H_1)}{2}\right)$, where BF_{10} represents the Bayes factor in favor of the alternative hypothesis and $BIC(H_1)$ and $BIC(H_0)$ denote the fit of the models with and without the fixed effect of interest, respectively. For interpretative convenience, we always display Bayes factors so that they reflect support for the favored hypothesis (i.e., BF_{10} when evidence favors the alternative hypothesis and BF_{01} when evidence favors the null hypothesis).

Sensitivity analyses. We conducted sensitivity analyses to determine the smallest effect size we were able to detect with 80% power (and $\alpha = 5\%$). We used the *simr* package (Green & Macleod, 2016) in R (R Core Team, 2021) to conduct sensitivity analyses for the main effects of interest (accuracy and accuracy awareness across all traits). The *simr* package does not include a function for conducting sensitivity analyses, but it does provide power estimates for fixed effects in multilevel regression models. We varied the effect of interest in our model and calculated power at each level. This allowed us to determine which effect size we were able to detect with 80% power.

Examining power for our multilevel regression model testing accuracy (i.e., the relation between perceivers personality ratings and targets' self-reported personality scores across all traits; see Results section) showed that we had 80% power to detect an effect of 0.068. Thus, we could detect a relation between perceivers' ratings and targets' self-reported scores where a one-point increase in targets' self-reported personality scores is associated with a 0.068-point

increase in perceivers' personality ratings. Next, we examined power for our model testing accuracy awareness (i.e., the interaction effect between targets' self-reported personality scores and perceivers confidence in their personality ratings across all traits; see Results section). This showed that we had 80% power to detect an effect of 0.016. Thus, we could detect a 0.016-point difference in the relation between perceivers ratings and targets' self-reported scores. In sum, our design had sufficient power to detect even low levels of accuracy and accuracy awareness.

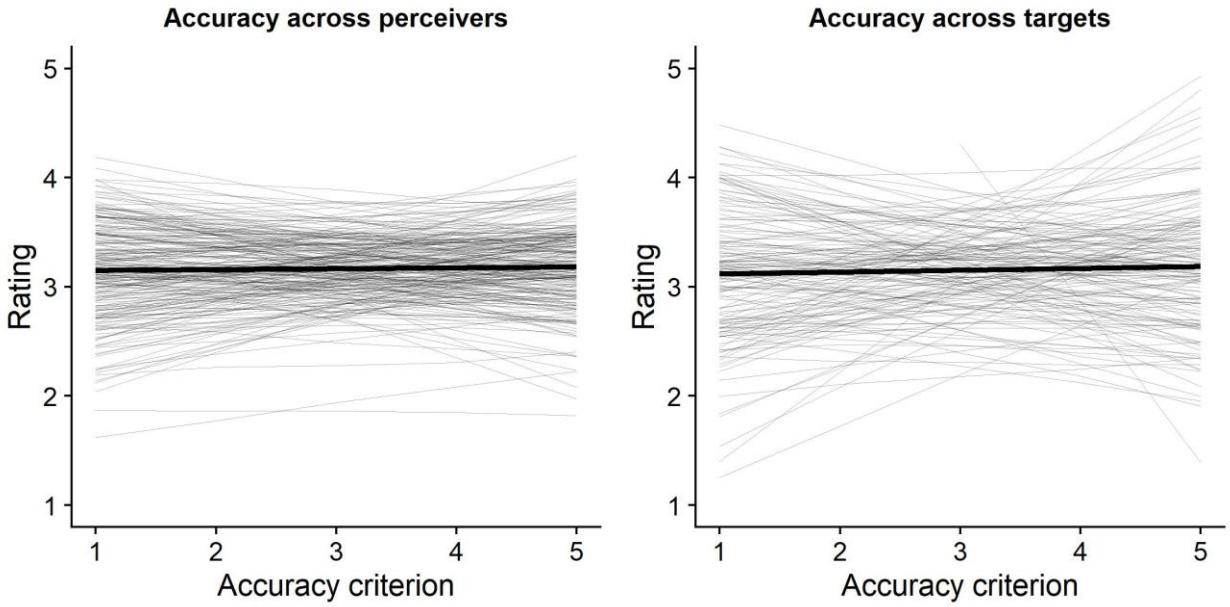
Results

Descriptive statistics. Perceivers' mean trait rating (averaged across all trials) ranged from 1.87 to 3.84 on our 5-point scale (M = 3.15, SD = 0.30). Perceivers' ratings also varied substantially on a trial-by-trial basis, with an average minimum rating of 1.25 (SD = 0.43) and an average maximum rating of 4.87 (SD = 0.33). Perceivers' mean confidence (averaged across all trials) ranged from 1.10 to 8.99 on our 9-point scale (M = 6.34, SD = 1.65). Perceivers' confidence also fluctuated on a trial-by-trial basis, with an average minimum confidence of 3.24 (SD = 1.61) and an average maximum confidence of 8.34 (SD = 0.99). These results show that both trait ratings and confidence were not uniformly low, high, or close to the midpoint of our scales, but they varied substantially within perceivers (between trials) and across perceivers (averaged across all trials).

Accuracy. First, we examined perceivers' mean trait accuracy across the five personality dimensions. We estimated a multilevel regression model in which perceivers' trait ratings were regressed on the accuracy criterion (i.e., targets' self-reported personality scores). This did not yield a significant effect and decisive evidence for the null hypothesis, b = 0.011, SE = 0.024, 95% CI [-0.035, 0.060], p = .648, BF_{01} > 1000. Thus, on average, across the five personality dimensions, perceivers did not show impression accuracy. There was significant variation in the slope of the association across targets, $\chi^2(2) = 84.59$, $p < .001$, but not across perceivers, $\chi^2(2) = 1.39$, p = .500 (see Figure 1). That is, although results suggested that some targets were judged significantly more accurately than others, they did not suggest that some judges were more accurate than others.

Figure 1

Variation in personality impression accuracy across perceivers and targets

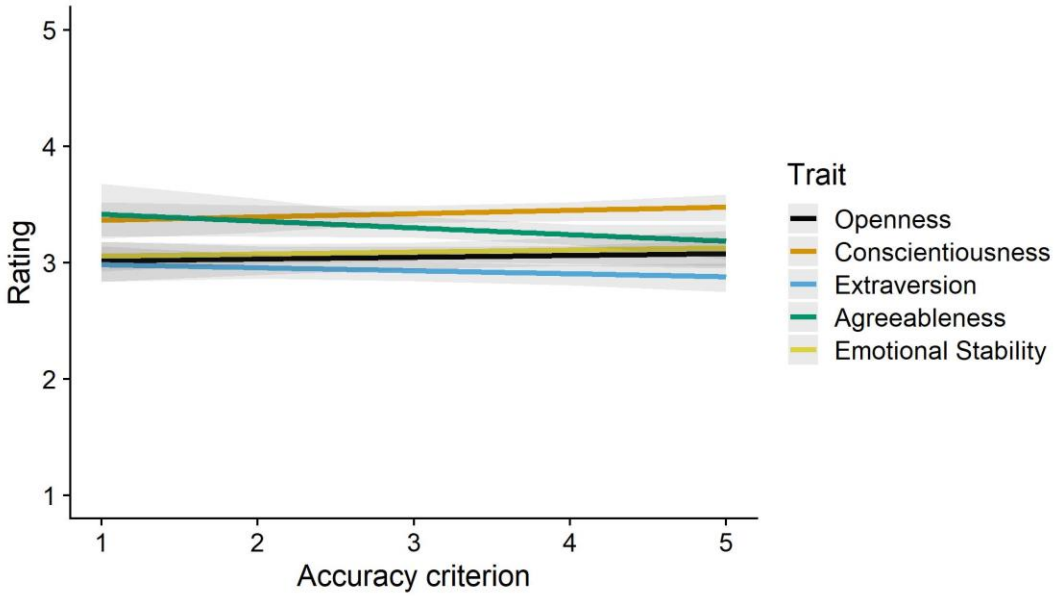


Note. The graphs visualize the association between personality impressions of perceivers and self-reported personality scores of targets (bold line), including variation in this association across perceivers (left) and across targets (right).

We also tested for trait accuracy per personality dimensions. Regressing trait ratings on the accuracy criterion (i.e., targets' self-reported personality), a variable indicating which personality dimensions was judged, and their interaction did not yield a significant interaction effect and decisive evidence in favor of the null hypothesis, $F(4, 189.1) = 1.14$, $p = .336$, $BF_{01} = 663.5$, suggesting that accuracy did not vary across the five dimensions. For each of the five dimensions, associations between trait ratings and targets' self-reported scores were not significant and there was decisive evidence in favor of the null hypothesis (openness: $b = -0.015$, $SE = 0.065$, 95% CI [-0.156, 0.121], $p = .816$, $BF_{01} = 398.3$; conscientiousness: $b = 0.044$, $SE = 0.048$, 95% CI [-0.045, 0.145], $p = .360$, $BF_{01} = 368.9$; extraversion: $b = -0.019$, $SE = 0.061$, 95% CI [-0.136, 0.100], $p = .750$, $BF_{01} = 420.8$; agreeableness: $b = -0.055$, $SE = 0.059$, 95% CI [-0.164, 0.048], $p = .348$, $BF_{01} = 294.2$; emotional stability: $b = 0.039$, $SE = 0.057$, 95% CI [-0.080, 0.137], $p = .491$, $BF_{01} = 375.4$; see Figure 2). Together, these results suggest that perceivers' impressions of targets' personality were not accurate.

Figure 2

The accuracy of perceivers' impressions for each personality dimension



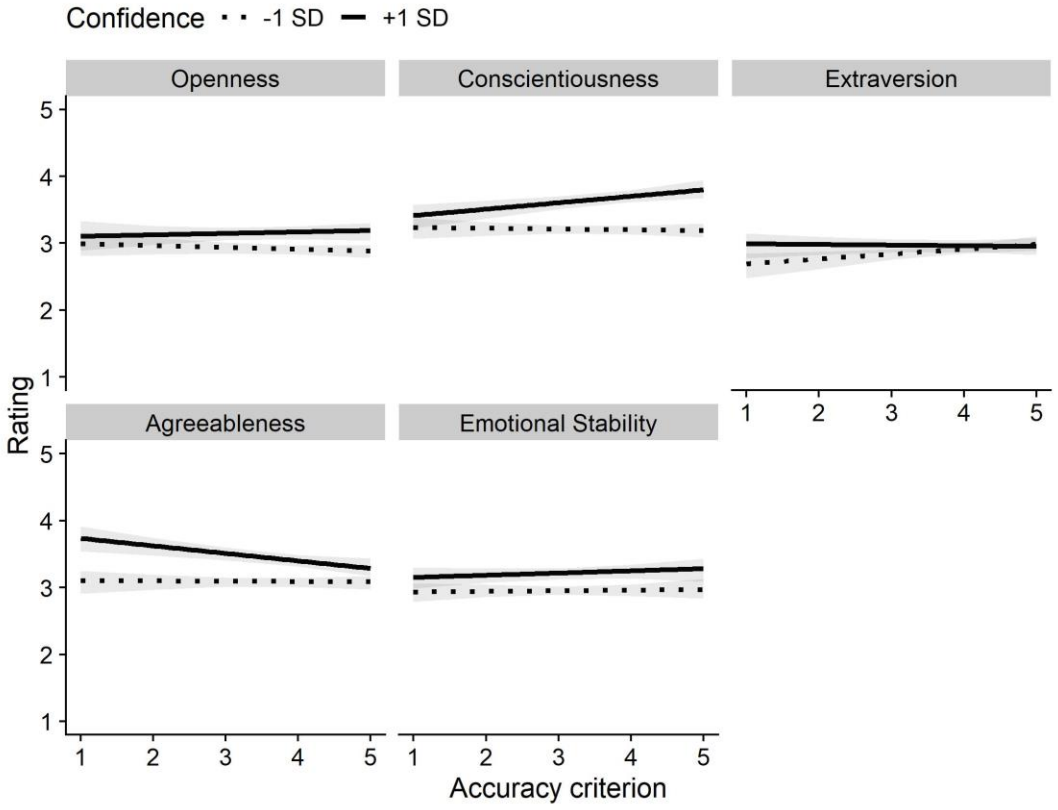
Accuracy awareness. Were perceivers aware of when their impressions were more or less accurate? We first examined this question by testing whether trial-level variation in confidence was associated with trial-level variation in accuracy (i.e., within-perceiver accuracy awareness). In other words, we examined whether there was a stronger association between perceivers' ratings and targets' self-reported scores when perceivers indicated higher levels of confidence. To test this, we estimated a multilevel regression model, in which we predicted perceivers' ratings with targets' self-reported scores, perceivers' confidence, and their interaction. Confidence scores were centered within perceivers (by subtracting each perceiver's average confidence across all trials) and then *z*-standardized. This analysis did not yield a significant interaction effect, but decisive support for the null hypothesis, b = -0.004, SE = 0.022, 95% CI [-0.051, 0.043], p = .863, BF_{01} > 1000, showing that accuracy was not higher when perceivers were more confident in the accuracy of their ratings.

We also tested for within-perceiver accuracy awareness per personality dimensions. We regressed perceivers' trait ratings on targets' self-reported scores, confidence, personality dimension, and their interactions. The three-way interaction was significant suggesting that there was variation in accuracy awareness across the five personality dimensions (although a Bayesian

analysis showed evidence in favor of the null hypothesis), $F(4, 2515) = 4.33$, $p = .002$, $BF_{01} > 1000$. At any rate, associations between perceivers' ratings and targets' self-reported scores were not moderated by confidence for any of the five dimensions (openness: $b = 0.022$, $SE = 0.030$, 95% CI [-0.038, 0.082], $p = .477$, $BF_{01} > 1000$; conscientiousness: $b = 0.046$, $SE = 0.024$, 95% CI [-0.009, 0.084], $p = .062$, $BF_{01} = 255.7$; extraversion: $b = -0.052$, $SE = 0.030$, 95% CI [-0.114, 0.012], $p = .090$, $BF_{01} = 239.5$; agreeableness: $b = -0.057$, $SE = 0.030$, 95% CI [-0.116, 0.004], $p = .059$, $BF_{01} = 150.9$; emotional stability: $b = 0.025$, $SE = 0.027$, 95% CI [-0.030, 0.080], $p = .355$, $BF_{01} = 666.1$; see Figure 3). Although three of the interaction effects were marginally significant, two were in the opposite direction (meaning that, if anything, perceivers were slightly *less* accurate when they were more confident). Bayesian analyses yielded decisive evidence in favor of the null hypothesis for all five dimensions. Thus, our results speak against the idea that perceivers have insight into the accuracy of their personality impressions.

Figure 3

Accuracy awareness when forming personality impressions



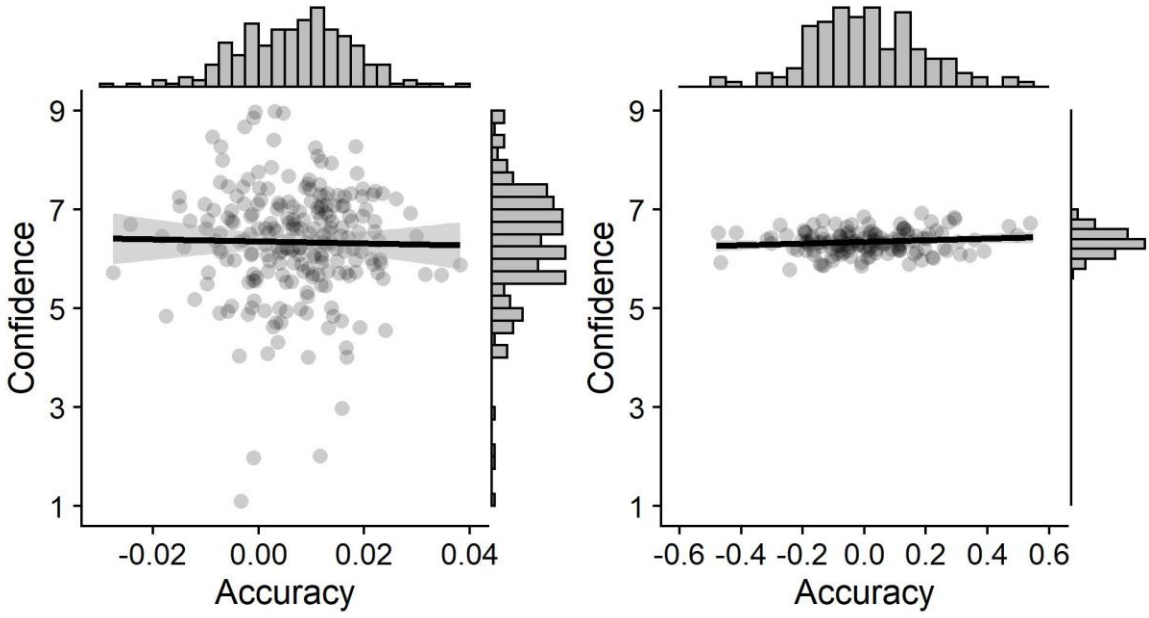
Note. The graph visualizes the relation between perceivers' impressions and targets' self-reported scores for each dimension when perceivers' confidence in the accuracy of their impression was low (i.e., one standard deviation below the mean; dotted lines) vs. high (i.e., one standard deviation above the mean; solid lines).

Next, we tested for between-perceiver accuracy awareness. We estimated a multilevel regression model in which perceivers' trait ratings were regressed on the accuracy criterion and we extracted the perceiver-specific slope. The perceiver-specific association between trait ratings and the accuracy criterion constituted our measure of perceiver-specific accuracy. Then, we calculated mean confidence scores for perceivers by averaging confidence ratings across all targets. There was no significant correlation between the average accuracy and average confidence of perceivers, with substantial evidence for the null hypothesis, $r(221) = -.02$, $p = .783$, $BF_{01} = 6.18$ (see Figure 4). That is, perceivers who were more confident were not more accurate.

We also tested for between-target accuracy awareness. From the model described above, we extracted the target-specific slope for the association between perceivers' ratings and the accuracy criterion, which constituted our measure of target-specific accuracy. We calculated mean confidence scores for targets (i.e., the average level of confidence for each specific target across all perceivers). There was no significant correlation (but only anecdotal evidence for the null hypothesis) between the average accuracy and confidence of targets (averaged across all perceivers), $r(138) = .13$, $p = .118$, $BF_{01} = 1.58$ (see Figure 4). That is, targets that were judged with greater confidence were not judged more accurately.

Figure 4

Accuracy awareness at the perceiver level (left) and at the target level (right). The left graph shows the relation between the average confidence and accuracy of perceivers (averaged across all targets). The right graph shows the relation between the average confidence and accuracy with which targets were judged (averaged across all raters)



Exploratory analyses. We also explored whether accuracy and accuracy awareness differed between male and female perceivers. Predicting perceivers' ratings with targets' self-reported scores, perceivers' gender, and their interaction, did not yield a significant interaction effect and decisive evidence for the null hypothesis, $b = 0.012$, $SE = 0.011$, 95% CI [-0.012, 0.034], $p = .269$, $BF_{01} > 1000$. In a similar vein, we did not find a significant three-way

interaction effect (with decisive evidence for the null hypothesis) between targets' self-reported scores, perceivers' confidence, and perceivers' gender, b = -0.003, SE = 0.005, 95% CI [-0.014, 0.007], p = .548, BF_{01} > 1000. Thus, we found no evidence that accuracy or accuracy awareness different between male and female perceivers.

Finally, we explored if there was a significant association between the aggregated judgments of all perceivers and our accuracy criterion. It is possible that targets' personality is better captured by the average judgment of multiple perceivers than by the judgments of individual perceivers (Ames et al., 2010; Naumann et al., 2009). For each target and trait, we calculated the average judgment across all perceivers. We did not find significant correlations between average trait judgments and targets' self-reported openness (with substantial evidence in favor of the null hypothesis), r(138) = -.01, p = .890, BF_{01} = 5.07, conscientiousness, r(138) = .09, p = .297, BF_{01} = 3.03, extraversion, r(138) = -.08, p = .367, BF_{01} = 3.46, agreeableness, r(138) = -.07, p = .420, BF_{01} = 3.74, or emotional stability, r(138) = .04, p = .667, BF_{01} = 4.68 (see the Supplemental Materials for a full correlation matrix between all self-reported and rated traits). Thus, we did not find that personality impressions are more accurate when they are averaged across many perceivers.

Discussion

Results of Study 1 point to a lack of accuracy and accuracy awareness when forming personality impressions based on others' facial appearance. Using larger samples of perceivers and targets than most previous studies, we found no significant associations between personality judgments based on a facial photograph and self-reported personality scores of targets. Bayesian analyses yielded strong support in favor of the null hypothesis. Similar null results were obtained when examining accuracy awareness. Our results do not support the idea that perceivers' confidence tracks their judgment accuracy across trials (i.e., an absence of within-perceiver accuracy awareness), or that perceivers who are generally more confident are also generally more accurate (between-perceiver accuracy awareness).

Study 2

In Study 2, we again tested perceivers' accuracy and accuracy awareness. We adapted the impression formation task to test if people are able to accurately judge others' personality under conditions that should make it easier for them to form accurate impressions. We showed perceivers pairs of faces and asked them to indicate which person scores higher on the trait in

question (rather than asking them to indicate a continuous rating for each target). This design simplifies the judgment process, providing perceivers with a clear reference for their judgment, as they only have to compare target A to target B, rather than to a larger, perhaps less clearly defined reference group. In Study 2, we focused on extraversion impressions due to resource constraints and as this is the dimension for which previous studies found the highest levels of accuracy (Borkenau et al., 2009; Penton-Voak et al., 2006).

We implemented three other changes in our study design compared to Study 1. First, we recruited raters from the United States and used an even larger sets of raters (n = 423) and targets (k = 1,260 unique target pairs with 280 unique targets). Second, we used facial photographs of both male and female targets and varied whether perceivers judged all-male, all-female, or mixed-gender pairs. This allowed us to explore if accuracy varies as a function of targets' gender. Previous research examining first impressions based on videos and brief interactions did not find differences in how accurately male and female targets were judged (Chan et al., 2011; Human et al., 2014). However, perceivers may be able to form more accurate impressions when making comparative judgments between men and women. Women generally score higher on extraversion (Costa et al., 2001; Feingold, 1994; Weisberg et al., 2011). Thus, target gender could be a valid cue and there is some evidence that perceivers rely on target gender when forming face-based impressions (Jaeger & Jones, 2021; Sutherland et al., 2015). Third, we adapted our confidence measure by letting perceivers bet coins on the accuracy of their judgment on each trial. Coins were doubled when perceivers' judgment was accurate and lost when it was inaccurate and we incentivized perceivers to maximize their total number of coins (perceivers did not receive feedback on the accuracy of their judgments or their accumulated number of coins). Perceivers also estimated how many of their impressions they expected to be correct. Comparing this estimate to their actual accuracy rate allowed us to test whether people are over- or underconfident in the accuracy of their impressions.

Methods

This study was preregistered (see https://osf.io/tr9zp/).

Participants. Simulation results suggest that trait ratings by approximately 20-25 unique raters produce relatively reliable average trait ratings of a target (Hehman et al., 2018). We therefore decided to recruit 420 participants, which would result in 30 unique ratings per target pair. Due to the randomization procedure with which perceivers were matched to target pairs, not

all target pairs had 30 unique ratings when we reached our planned sample size. We therefore continued to recruit participants until all target pairs had been rated at least 30 times, leading to a slightly larger sample size than preregistered. In total, we recruited 424 U.S. American workers from Amazon Mechanical Turk (Amir et al., 2012; Paolacci & Chandler, 2014) who completed the study in return for $1 and three chances to receive a $25 bonus payment. It took participants approximately 8 minutes to complete the study. In line with our preregistered exclusion criteria, data from one participant (0.24%) who indicated that they completed the study on a cell phone were excluded, leaving a final sample of 423 participants (M_{age} = 38.4 years, SD_{age} = 11.0; 44.21% female, 54.61% male, 1.18% other).

Stimuli. We used the same 140 facial photographs of female students from a German University as in Study 1. We also used a set of 163 facial photographs of male students from the same population (18-34 years old). From this set, we selected the first 140 targets in order to balance the number of male and female targets. All targets were photographed in front of a white background and showed a neutral facial expression (Kordsmeyer et al., 2018). Targets' personality was assessed with the German version of the 42-item Big Five Inventory (Lang et al., 2001). Extraversion scores showed good internal reliability, $\alpha = .87$.

The photographs were displayed in pairs. We first created all unique pairs based on our sample of 280 faces (k = 39,080). Target pairs in which both individuals had the same personality score were discarded (k = 37,035 remaining). From this set, we randomly sampled 1,260 pairs with the following restrictions: each target was included 9 times—6 times paired with a target from the same sex and 3 times paired with a target from the other sex. Thus, our final stimulus set contained 1,260 target pairs: 420 all-female pairs, 420 all-male pairs, and 420 mixed-gender pairs.

Procedure. Perceivers completed 90 trials at a self-paced rate. On each trial, perceivers saw a randomly drawn target pair. This means that it was possible for perceivers to see the same target more than once. However, given that each target was only shown in 9 out of 1,260 target pairs, it was unlikely that perceivers saw the same face many times (i.e., a 6% chance to see the same face twice, a 0.4% chance to see the same face three times). Perceivers indicated their extraversion impressions by selecting the person that they think is more extraverted. We measured perceivers' confidence after each rating. Perceivers received 10 coins that they could either keep or bet on the accuracy of their rating. When perceivers bet the coins and their rating

was correct, the coins were doubled (i.e., they received 20 coins). When perceivers bet the coins and their rating was incorrect, the coins were lost (i.e., they received 0 coins). When perceivers decided not to bet, they received 10 coins. Thus, to maximize their total point count, perceivers had to bet the coins when they were more confident in the accuracy of their rating and keep the coins when they were less confident. Participants did not receive feedback on whether their judgments were accurate and they did not see their point coin (from which they could infer whether their previous rating was accurate).

We also measured perceivers' confidence by asking them to predict their overall performance. After completing all trials, perceivers indicated how many target pairs they thought they had judged correctly on a scale that ranged from 0% to 100%. We reminded perceivers that approximately half of their ratings should be accurate by chance alone. This measure allowed us to test whether perceivers were over- or underconfident in the accuracy of their impressions, by comparing perceivers' expected and actual accuracy.

We again incentivized perceivers' accuracy and accuracy awareness. Perceivers were informed that the person with the most accurate ratings (i.e., the person with the highest number of correct extraversion ratings), the person with the highest accuracy awareness (i.e., the person who accumulated the most coins after 90 trials), and one person who correctly guessed their percentage of accurate ratings would each be rewarded with a $25 bonus payment.

Analysis strategy. Perceivers' ratings were coded as 1 when they were accurate and as 0 when they were inaccurate, depending on whether the target that was judged to be more extraverted actually had a higher extraversion score. We estimated generalized multilevel regression models with random intercepts for perceivers and target pairs to model variation in accuracy (when testing for accuracy) and confidence (when testing for accuracy awareness). When testing for accuracy awareness, we also included random slopes per perceiver and target pair to model variation in accuracy awareness. For all primary tests, we report the results of frequentist and Bayesian analyses. We explored the robustness of our results by implementing different priors (see Supplemental Materials).

Sensitivity analyses. We used the *simr* package (Green & Macleod, 2016) in R (R Core Team, 2021) to conduct sensitivity analyses for the main effects of interest (testing the accuracy and accuracy awareness of extraversion judgments). Examining power for our model testing accuracy (i.e., the percentage of times perceivers made an accurate judgment compared against

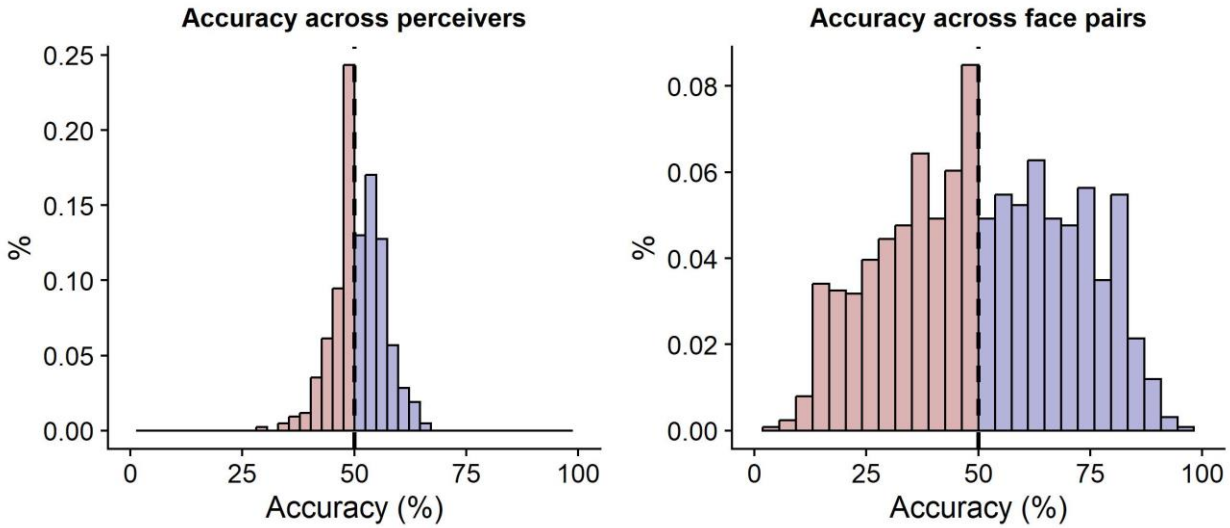
chance) showed that we had 80% power to detect an accuracy level of 51.82%. Next, we examined power for our model testing accuracy awareness (i.e., the relation between betting behavior and accuracy). This showed that we had 80% power to detect an odds ratio of 1.09. Thus, when comparing accuracy when people were betting (vs. not betting) on their judgment, we could detect a change in accuracy from, for example, 50.00% to 52.16%. Thus, our design had sufficient power to detect even low levels of accuracy and accuracy awareness.

Results

Accuracy. Ratings were accurate 51.10% of the time. We tested whether ratings were accurate significantly more often than expected by chance (i.e., 50%) by examining the intercept in a multilevel regression model with rating accuracy as the outcome and random intercepts per perceiver and target pair. This yielded an intercept that was just significant, $b = 0.051$, $SE = 0.026$, $OR = 1.05$, 95% CI [1.00, 1.11], $p = .049$, $BF_{01} = 28.31$. However, a Bayesian analysis indicated strong support in favor of the null hypothesis (i.e., accuracy not being different from 50%). There was significant variation in accuracy across targets, $\chi^2(2) = 3242$, $p < .001$, and across perceivers, $\chi^2(2) = 4.83$, $p = .028$ (see Figure 5).

Figure 5

The distribution of accuracy in extraversion impressions across perceivers (left) and across face pairs (right)



Note. Dotted lines denote chance-level accuracy (i.e., 50%). Perceivers whose average accuracy across all trials was larger than 50% and face pairs that were judged with more than 50% accuracy (averaged across all raters) are displayed in blue. Perceivers whose average accuracy across all trials was smaller than 50% and face pairs that were judged with less than 50% accuracy (averaged across all raters) are displayed in red.

Accuracy awareness. Next, we examined perceivers' accuracy awareness by analyzing their betting behavior. Perceivers bet on 56.00% of all trials, with 41 perceivers (9.69%) always betting and 22 perceivers (5.20%) never betting. Perceivers were incentivized to bet (vs. not bet) when they thought that their rating was accurate (vs. inaccurate), as this would lead to the highest payoffs. We realized that in this context, the behavior of perceivers who always or never bet is difficult to interpret. Both strategies lead to the same earnings if perceivers believe that their ratings are not accurate at all (50% accuracy). For perceivers who always bet, betting on a given trial is not a good measure of confidence as it could reflect both extreme confidence (expected accuracy of 100%) or the complete lack thereof (expected accuracy of 0%). We therefore decided to exclude invariant bettors (n = 63, 14.89 %) from all analyses of betting decisions. As this exclusion criterion was not included in our preregistration, we also report analyses that included invariant bettors, which produced qualitatively similar findings, in the Supplemental Materials.

To test for within-perceiver accuracy awareness, we estimated a multilevel regression model, in which we predicted betting behavior (0 = did not bet, 1 = did bet) with rating accuracy (0 = inaccurate rating, 1 = accurate rating). This did not yield a significant effect and decisive evidence for the null hypothesis, b = -0.010, SE = 0.033, OR = 0.99, 95% CI [0.93, 1.11], p = .765, BF_{01} = 172.3. In other words, perceivers' extraversion impressions were not more accurate when perceivers were more confident in them.

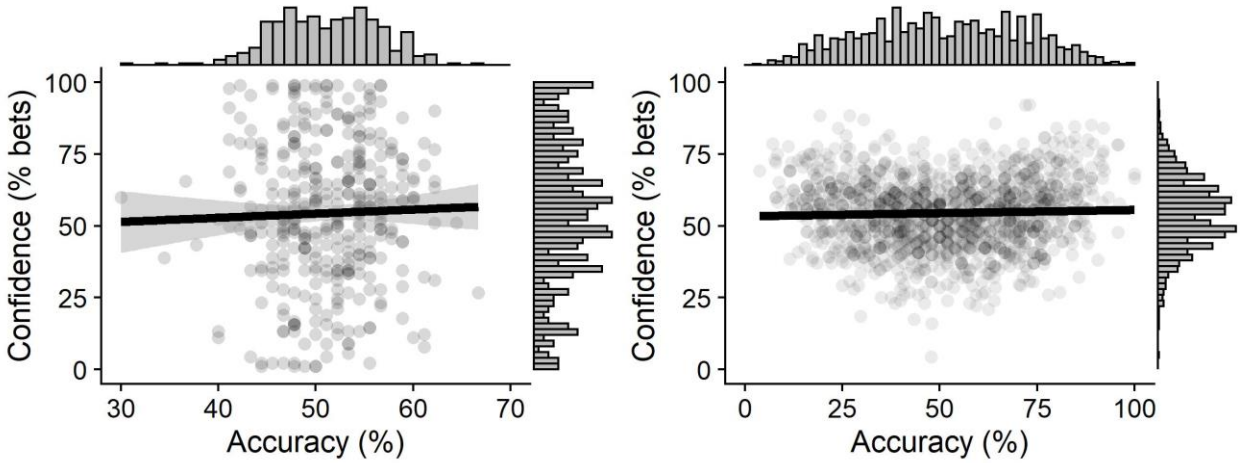
We also tested for between-perceiver accuracy awareness. For each perceiver, we calculated a confidence score (their betting frequency across all trials) and an accuracy score (how often their rating was accurate across trials). There was no significant correlation between confidence and accuracy with substantial evidence in favor of the null hypothesis, $r(358) = .03$, p = .566, BF_{01} = 6.91 (see Figure 6). That is, perceivers who were on average more confident were not more accurate.

We tested for between-target accuracy awareness by calculating a confidence score (how often perceivers bet when judging a given face pair) and an accuracy score (how often a give face pair was judged accurately across all perceivers) at the target level. The correlation between confidence and accuracy was not significant with substantial evidence in favor of the null hypothesis, $r(1258) = .04$, p = .190, BF_{01} = 6.44 (see Figure 6). That is, targets that were on average judged with greater confidence were not judged more accurately.[5]

[5] An additional exploratory analysis suggested the presence of a quadratic effect (see Supplemental Materials).

Figure 6

The relation between average confidence and average accuracy of extraversion impressions at the perceiver level (left) and at the target level (right)



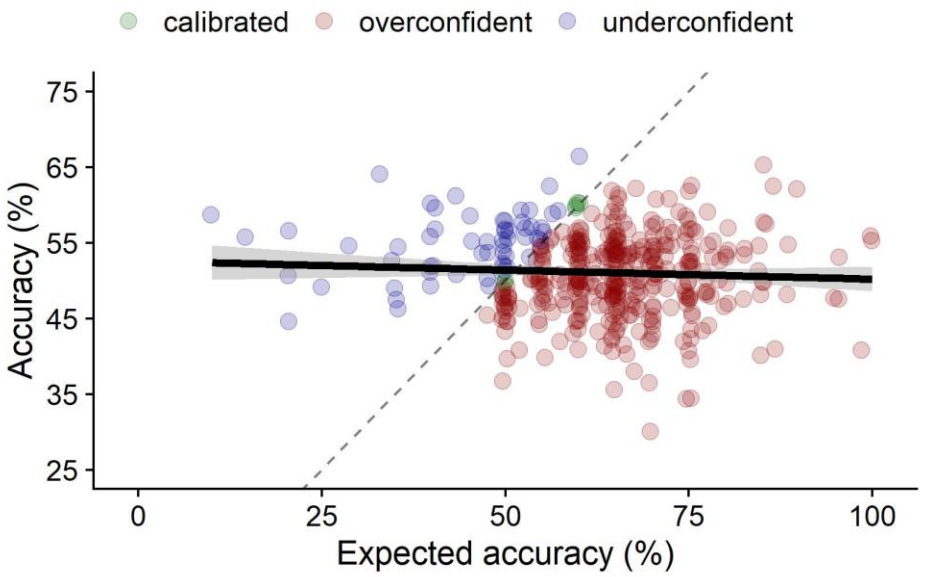
Note. The left graph shows the relation between the percentage of times perceivers bet on the accuracy of their ratings and perceivers' accuracy (averaged across all targets). The right graph shows the relation between the percentage of times face pairs were bet on and the accuracy with which face pairs were judged (averaged across all raters).

Despite this apparent lack of accuracy awareness, perceivers overall winnings were slightly higher than expected by chance. A person who bets randomly (thus winning on half of all trials) has an expected return of 10 coins per trial and would therefore accumulate 900 coins. On average, perceivers accumulated 912.6 coins (*SD* = 70.08), which was significantly different from 900 (with strong evidence for the alternative hypothesis), $t(359) = 3.42$, $p < .001$, $d = 0.18$, $BF_{10} = 17.96$.

Finally, we analyzed perceivers' overall confidence in their performance. On average, perceivers expected 63.09% of their extraversions ratings to be accurate (*SD* = 12.47%). The correlation between expected and actual accuracy was not significant with substantial evidence in favor of the null hypothesis, $r(421) = -.06$, $p = .256$, $BF_{01} = 4.65$, again suggesting that perceivers were not aware of their actual accuracy (see Figure 7). Moreover, expected accuracy was significantly higher (with decisive evidence in favor of the alternative hypothesis) than actual accuracy, $t(422) = 17.77$, $p < .001$, $d = 0.86$, $BF_{10} > 1000$, showing that perceivers tended to be overconfident in the accuracy of their impressions. Substantially more perceivers overestimated (84.40%), rather than underestimated (14.18%) their accuracy (1.41% provided accurate estimations).

Figure 7

The relation between perceivers' expected and actual accuracy of extraversion impressions



Note. The dashed line represents perfect accuracy (i.e., a correlation between expected and actual accuracy of 1). Deviations to the left of the line signify underestimations of accuracy while deviations to the right signify overestimations over accuracy. Thus, Green dots represent perceivers that correctly guessed their accuracy, blue dots represent perceivers that were underconfident, and red dots represent perceivers that were overconfident.

Exploratory analyses. We explored whether differences in extraversion scores of targets influenced the accuracy of impressions. If differences in extraversion are reflected in facial features and are therefore readable by perceivers to some extent, then perceivers should be able to provide more accurate judgments when two targets differ a lot (vs. a little) on extraversion. Across all face pairs, the absolute difference in extraversion scores ranged from 0.12 to 3.12 points on our five-point scale (M = 0.82, SD = 0.58). Regressing the accuracy of perceivers' judgments on the extraversion difference of targets did not produce a significant effect and decisive evidence in favor of the null hypothesis, b = -0.031, SE = 0.043, OR = 0.97, 95% CI [0.89, 1.05], p = .467, BF_{01} = 149.8. Thus, extraversion impressions were not more accurate when targets actually differed more on extraversion.

We also explored whether accuracy or accuracy awareness varied as a function of target gender (all-male vs. all-female vs. mixed-gender pairs) or perceiver gender (male vs. female),

but found no significant effects and decisive evidence in favor of the null hypotheses (all BF_{01} > 1000; see Supplemental Materials for full results).

Finally, we again tested if there was a significant association between the aggregate judgments of all perceivers and our accuracy criterion. For each target pair, we calculated the average judgment across all perceivers (i.e., the percentage of perceivers that judged a specific target in a given pair to score higher on extraversion). The point-biserial correlation between the percentage of perceivers that judged a certain target to be relatively more extraverted and whether this target was actually more extraverted than the other target in the pair was not significant (with substantial evidence in favor of the null hypothesis), $r(1258) = .06$, $p = .067$, $BF_{01} = 6.58$.

Discussion

In Study 2, we tested perceivers' accuracy and accuracy awareness when guessing which of two targets scores higher on extraversion. Accuracy was just significantly above chance (51.1%, $p = .049$), and a Bayesian analysis indicated strong support in favor of the null hypothesis (i.e., accuracy not being different from 50%). Thus, the present results provide only very weak support for the idea that extraversion impressions based on others' facial appearance are accurate. The evidence for (a lack of) accuracy awareness was less ambiguous. As in Study 1, we did not find evidence for within-perceiver or between-perceiver accuracy awareness. Bayesian analyses showed strong support for chance-level accuracy awareness. We also found that the vast majority of perceivers (84.4%) were overconfident in the accuracy of their extraversion impressions.

General Discussion

People form snap judgments of other's personality based on their facial appearance and these impressions influence many consequential decisions (Todorov, Olivola, et al., 2015). Here, we provide novel evidence on the accuracy of personality impressions—which has been extensively studied, but with inconsistent results (Borkenau et al., 2009; A. L. Jones et al., 2012; Penton-Voak et al., 2006)—and people's accuracy awareness—which has received little attention despite its theoretical and practical importance (Ames et al., 2010).

Overall, our findings suggest that personality impressions from faces do not reflect targets' actual personality, and that people are not aware of their (in-)accuracy. In other words, our findings suggest that perceivers lack accuracy and accuracy awareness when forming

personality impressions from faces. These conclusions are supported by Bayesian analyses, which yielded (often very strong) evidence in favor of the hypothesis that levels of accuracy and accuracy awareness are at chance level. Only Study 2 yielded an estimate of 51.10% accuracy for extraversion impressions, which was just significantly higher than chance (p = .049). Although we leave the interpretation of this result open to the reader, we do not consider it convincing evidence in favor of accuracy, especially because a Bayesian analysis indicated strong support in favor of the null hypothesis. Perceivers showed relatively low levels of consensus in their judgments, which also speaks against the idea that their judgments reflect a target's underlying trait.

Chance-level accuracy and accuracy awareness was obtained (a) for all dimensions of the Big Five, (b) for continuous and binary judgments (we only examined binary judgments of extraversion in Study 2), (c) with perceivers from the Netherlands and the United States, and (d) irrespective of perceiver and target gender. Accuracy and accuracy awareness were not above chance even though we employed judgment tasks that incentivized perceivers to give accurate personality judgments and accurate estimates of their judgment accuracy, and even though we relied on considerably larger samples of raters and targets compared to previous studies, meaning that we had sufficient power to detect even low levels of accuracy and accuracy awareness. We found that a perceiver's judgments were not more accurate when the perceiver was more (vs. less) confident in them (a lack of *within*-perceiver accuracy awareness) and that perceivers who were on average more (vs. less) confident were on average not more accurate (a lack of *between*-perceiver accuracy awareness). In fact, comparing perceivers' estimated accuracy with their actual accuracy in Study 2 showed that they tended to be overconfident. On average, perceivers expected their judgments to be correct 63% of the time, even though they were only correct 51% of the time. This was not due to a few perceivers being much more confident. We found that 84% of perceivers overestimated their accuracy.

Prior evidence on the accuracy of personality impressions from faces is mixed. Almost every study found a different pattern of results regarding which personality dimension can or cannot be inferred with some level of accuracy based on a facial photograph. Perhaps the most consistent evidence in favor of accuracy was found for impressions of extraversion (Borkenau et al., 2009; Kramer & Ward, 2010; Naumann et al., 2009; Penton-Voak et al., 2006; Satchell et al., 2019). Still, results from several studies did not yield support for accurate extraversion

impressions (Ames et al., 2010; A. L. Jones et al., 2012; Shevlin et al., 2003). Our results are in line with these latter findings.

To situate the current findings within the existing literature, two points are important to note. First, it is plausible that inconsistencies with prior results are due to methodological differences between studies. Prior studies relied on different types of face stimuli, including self-selected profile photos (Satchell et al., 2019), facial photographs taken under standardized conditions in the lab (Kramer & Ward, 2010), and composite images (Penton-Voak et al., 2006). Richer stimuli including more information about the displayed person, such as profile photos, may lead to more accurate personality impressions than the more standardized images that were used in the present studies. Similarly, accuracy may be higher when targets display spontaneous facial expressions when photographed (see, for example, Borkenau et al., 2009). While these hypotheses are plausible (e.g., Borkenau & Liebler, 1992; Kenny & West, 2008), they are not sufficient to explain the pattern of results in published findings. Various investigations that used standardized photographs of resting faces against a neutral background did report evidence for accurate personality judgments (Naumann et al., 2009; Nestler et al., 2012; Penton-Voak et al., 2006). Most studies that did not restrict targets' facial expression or used profile photos from social media platforms found evidence for accurate extraversion impressions, representing perhaps the most consistent pattern of results in previous findings (Borkenau et al., 2009; Borkenau & Liebler, 1992; Satchell et al., 2018; Stopfer et al., 2014). Yet, evidence on impression accuracy for the other Big Five dimensions was inconsistent and one study analyzing profile photos found non-significant levels of accuracy across all five dimensions (Qiu et al., 2015). More systematic investigations are needed to disentangle how these different factors, such as emotional expressivity, influence accuracy for different dimensions.

Second, it is plausible that inconsistencies in prior results were due to methodological shortcomings that can lead to false positive or false negative results. Most studies examined accuracy for all Big Five dimensions, sometimes testing accuracy separately for male and female targets (Little & Perrett, 2007; Penton-Voak et al., 2006). Without correction for multiple testing, this procedure can inflate the rate of false positive results (Simmons et al., 2011). Positive findings are also more likely to be reported and published (Francis, 2014; L. K. John et al., 2012). In other words, it is plausible that additional studies that did not find evidence for accuracy exist, but were never published. The majority of prior studies also relied on relatively

small samples of raters and targets (Ames et al., 2010; Shevlin et al., 2003), which limits statistical power and can lead to false negative results. These considerations were key motivators behind the current studies, in which we relied on much larger samples of perceivers and targets to provide a more reliable test of accuracy in personality impressions.

Theoretical and Practical Implications

Theories of social perception that aim to explain why and how people form personality impressions based on others' facial appearance can be grouped into two broad categories. Some theoretical accounts posit that people rapidly form and rely on face-based personality impressions because some facial cues (e.g., facial width-to-height ratio) are valid indicators of a targets' personality and reliance on these facial cues allows perceivers to make accurate judgments (Carré et al., 2009; Carré & McCormick, 2008; Stirrat & Perrett, 2010). Other accounts (e.g., overgeneralization theory) posit that face-based personality impressions are not necessarily accurate because they are byproducts of otherwise adaptive social-cognitive mechanisms, such as a heightened sensitivity to detect and rely on facial expressions when judging others (Todorov, Olivola, et al., 2015; Zebrowitz, 2012). The current results lend support to the latter view. Even though our studies were powered to detect even low levels of accuracy and perceivers were incentivized to form accurate impressions, our findings suggest that perceivers facial impressions are not accurate.

People rely on facial impressions when making many consequential decisions, including voting, sentencing, and hiring decisions (Olivola et al., 2014). The current findings suggest that this widespread reliance on personality impressions from faces is problematic for two reasons. We find that personality impressions from faces are not accurate. In other words, it is likely that people treat others differently because they falsely attribute certain traits to them based on their appearance. This does not necessarily imply that people should never rely on their impressions. Selective reliance would be justified if people can discriminate between situations in which their impressions are more accurate and can be relied upon, and instances in which their judgments are inaccurate and should not be relied upon. However, our findings also suggest that people lack such insight.

Limitations and Future Directions

Although the current studies provide consistent evidence for a lack of accuracy and accuracy awareness, more work on this topic is needed. In both studies, we employed

photographs of German targets and examined personality impressions of perceivers from Western societies. Future studies could test the robustness of our results using more diverse samples or targets and perceivers (see, for example, B. C. Jones et al., 2021). More work is also needed to explore perceivers' accuracy and accuracy awareness when making other types of trait judgments. For example, it is still unclear whether trustworthiness impressions from faces are accurate (Foo et al., 2022; Jaeger, Oud, et al., 2022; Siuda et al., 2022).

Future studies should also test accuracy and accuracy awareness for different types of stimuli. When judging others' personality, people rely on their facial appearance but also many other cues (Alaei & Rule, 2016; Back & Nestler, 2016). Previous work has examined judgment accuracy for dozens of stimuli and situations, including minimal static stimuli such as eyes (Bjornsdottir et al., 2017) or shoes (Gillath et al., 2012), richer static stimuli such as face images (Nestler et al., 2012) or social media websites (Van De Ven et al., 2017), and much richer, dynamic situations such as unstructured face-to-face interactions (Biesanz et al., 2011). In the present studies, we focused on facial photographs because an extensive literature shows that people rely on facial appearance to form trait impressions and that these impressions influence various decisions (Todorov, Olivola, et al., 2015). Although voting, legal sentencing, personnel selection, and virtually every other important decision process is the product of many factors and considerations, there is robust evidence that facial impressions contributes to many of these decisions (Graham et al., 2016; Jaeger, Sleegers, et al., 2019; Olivola et al., 2014). Many people also believe in their ability to judge a person's character with some accuracy based on their facial appearance (Jaeger, Evans, et al., 2022; Suzuki et al., 2017) and facial impressions influence decision-making even when people have access to other, more valid cues (Olivola & Todorov, 2010b; Rezlescu et al., 2012) or when they are instructed to ignore facial appearance (Hassin & Trope, 2000; Jaeger et al., 2020). These observations raise the question of whether perceivers are actually able to judge others' personality based on their facial appearance with some accuracy, which we aimed to address here.

Extending previous work on this topic (e.g., Borkenau et al., 2009; Naumann et al., 2009; Nestler et al., 2012), we found that participants in our studies were not accurate and also ignorant of their inaccuracy. However, results may be different when impressions are based on stimuli that are richer than a facial photograph. For example, many studies have replicated the finding the personality judgments, especially of others' extraversion, after brief face-to-face interactions

are somewhat accurate (Biesanz et al., 2011; Borkenau & Liebler, 1992; for a review, see Back & Nestler, 2016). Thus, although the present results suggest that people's reliance on facial appearance when judging others' personality is detrimental for their judgment accuracy, this does not mean that their overall judgments, based on all available cues that they may rely on in a given situation, will always be inaccurate. Testing whether perceivers can form accurate impressions based on a certain cue provides valuable insights, especially if there is strong evidence that perceivers rely on this cue in everyday life, even when they may have access to other cues. However, it is also important to study impression accuracy under less controlled conditions in which perceiver's have access to a host of cues. For example, previous work has focused on impression accuracy based on brief face-to-face encounters (Biesanz et al., 2011) and social media profiles (Van De Ven et al., 2017; Vazire & Gosling, 2004). We see these as complimentary approaches.

One limitation of the current studies was their exclusive reliance on self-reports to assess targets' personality. Although self-reports have been shown to reliably predict a variety of important outcomes (Roberts et al., 2007; Soto, 2019), they are also subject to socially desirable responding and other biases, which is why a combination of self-reports and observer ratings is widely considered as the gold standard for accuracy criteria (Funder, 1995; Vazire & Gosling, 2004). A salient concern is that, although perceivers may be capable of judging others "true" personality with some accuracy based on a facial photograph, this accuracy does not emerge because the benchmark with which accuracy is assessed, target's self-reported personality, reflects targets' projected or idealized self rather than their true self. This interference (and a resulting lack of impression accuracy) should be observed for traits that respondents are particularly motivated to project, such as agreeableness (Graziano & Tobin, 2002; Paulhus et al., 1995).

Although we cannot rule out that this had some influence on the present results, we deem it unlikely that the impact on accuracy estimates was large. We found no evidence for accuracy even for the trait dimensions that should be less affected by social desirability bias (e.g., openness, extraversion). Most previous studies also relied on self-reports and some found evidence for accurate impressions of agreeableness (A. L. Jones et al., 2012; Nestler et al., 2012). Other studies relied on informant ratings (Ames et al., 2010) or composite scores of self-reports and informant ratings, which should limit the impact of socially desirable responding on

accuracy (Alper et al., 2020; Naumann et al., 2009; Shevlin et al., 2003). However, these studies did not yield stronger evidence in favor of impression accuracy and the pattern of results across these studies was as inconsistent as the results of studies that only relied on self-reports. Thus, it seems unlikely that the absence of accuracy in the current or previous studies, or the generally inconsistent pattern of findings in the literature can be explained by social desirability bias in self-reports.

Similar to previous work (Borkenau et al., 2009, Little & Perrett, 2007; Satchell et al., 2019, see also Alper et al., 2020), we assessed personality impressions using a single item with a description of the relevant trait dimension, whereas targets' personality was assessed with a 44-item questionnaire. Different interpretations of the measures by targets and perceivers may artificially suppress observed relations between self-reported and rated personality. However, it is also not obvious that having perceivers rate targets on the same Big Five inventory is the preferred alternative. In everyday life, people likely judge others along a few, relatively broad dimensions when they have to base their judgments on superficial cues such as a target's appearance (Oosterhof & Todorov, 2008; Sutherland et al., 2017). Thus, ratings of broad dimensions (e.g., this person is agreeable), rather than specific behaviors and tendencies (e.g., this person has few artistic interests, this person perseveres until the task is finished), may better capture how impressions are formed in everyday life.

Although it is sometimes not clear which design choices are optimal, we see improvements in measurement practices as an important next step for the first impression accuracy literature. This applies especially to the large literature on the accuracy of trustworthiness impressions, which has also produced many inconsistent findings (Foo et al., 2022; Jaeger, Oud, et al., 2022; Siuda et al., 2022). Studies often relied on small target samples and a recent meta-analysis found that a third of all effect sizes were based on the same three stimulus sets (and should therefore not be treated as independent estimates; Foo et al., 2022). Targets' behavior in a trust game is usually taken as the accuracy criterion, but recent work has questioned the validity of economic games for capturing individual differences in social preferences (Banerjee et al., 2021; Galizzi & Navarro-Martinez, 2019), especially when they are administered once (X. Wang & Navarro-Martinez, 2023). Future tests should rely on large samples of targets and raters and improved measures of trustworthiness (for an example, see X. Wang & Navarro-Martinez, 2023).

Future studies could also examine potential moderators of accuracy and accuracy awareness. Although we found that personality impressions from faces were not accurate on average, it is possible that impressions are accurate for some types of perceivers or some types of targets. For example, work on the accuracy of personality judgments in richer social contexts (e.g., brief face-to-face interactions) has identified a number of judge characteristics (Biesanz & Human, 2010; Capozzi et al., 2020) and target characteristics (Human et al., 2014; Human & Biesanz, 2013; Kerr et al., 2020) that moderate accuracy. This is also highlighted by the Realistic Accuracy Model (Funder, 1995), one of the most influential models of personality judgments, which posits that there are "good judges", perceives who are consistently more accurate in their judgments, and "good targets", individuals who are consistently judged more accurately. In line with previous work (Biesanz, 2010), we found more variation in accuracy across targets than across perceivers in both studies (additional exploratory analyses of variation in accuracy are reported in the Supplemental Materials). This suggests that explorations of potential moderators of judgment accuracy may be more successful when they focus on target characteristics rather than perceiver characteristics.

Yet, it is also plausible that the same does not apply to face-based personality judgments. The Realistic Accuracy Model identifies necessary conditions for accuracy and it is plausible that one or more conditions are not met when perceivers judge others only based on their facial appearance. It is plausible that a person's face does not contain any relevant cues to their personality, or that perceivers rely on the wrong cues when forming personality impressions. Additional work is needed to examine these open questions, for example, by assessing targets' personality, their facial features (e.g., attractiveness, babyfacedness, sexual dimorphism), and perceivers' personality impressions and confidence. A lens model approach (Brunswik, 1956; Nestler & Back, 2013) would allow researchers to test whether valid facial cues are available (i.e., relations between personality scores and specific facial cues) and which cues perceivers rely on when forming personality impressions (i.e., relations between specific facial cues and perceivers' judgments). This could also provide novel insights into how different facial features influence variation in impression accuracy and perceivers' (over-)confidence in their impressions. For example, exploratory analyses (reported in the Supplemental Materials) showed that participants in Study 2 were more confident when judging pairs of female targets than when judging pairs of male targets or mixed-gender pairs. Perceivers were more confidence when

forming extraversion impression than when forming emotional stability impressions, but these differences were very small.

In line with previous work that examined the accuracy of face-based trait impression (Ames et al., 2010; Borkenau et al., 2009; Naumann et al., 2009; Satchell et al., 2019), our analyses focused on how accurately perceivers can distinguish different targets on a given trait, usually referred to as trait accuracy (Biesanz, 2010; Hall et al., 2018). We also examined, and found no evidence for, aggregate observer accuracy (Ames et al., 2010; Borkenau et al., 2009; Naumann et al., 2009). Additional indicators of accuracy have been distinguished in the literature. For example, previous studies have examined the extent to which perceivers' trait judgments reflect characteristics of the average person (i.e., normative accuracy) and the extent to which perceivers can judge the relative level of different traits within a given target (i.e. profile accuracy; Hall et al., 2018; Krzyzaniak et al., 2019; Naumann et al., 2009). In the majority of these studies, perceivers judged targets after engaging in a brief face-to-face interaction or after watching a video of the target. Future studies could apply similar analytic approaches (Biesanz, 2010) to better understand sources of accuracy and bias in facial impressions.

Data Accessibility Statement

All data, analysis scripts, and preregistration documents are available at the Open Science Framework (https://osf.io/tr9zp/). We report how our sample sizes were determined and all data exclusions and measures for each study.

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

Adams, R. B., Nelson, A. J., Soto, J. A., Hess, U., & Kleck, R. E. (2012). Emotion in the neutral face: A mechanism for impression formation? *Cognition & Emotion*, *26*(3), 431–441. https://doi.org/10.1080/02699931.2012.666502

Alaei, R., & Rule, N. O. (2016). Accuracy of perceiving social attributes. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The Social Psychology of Perceiving Others Accurately* (pp. 125–142). Cambridge University Press. https://doi.org/10.1017/cbo9781316181959.006

Alper, S., Bayrak, F., & Yilmaz, O. (2020). All the Dark Triad and some of the Big Five traits are visible in the face. *Personality and Individual Differences*, *168*, 110350. https://doi.org/10.1016/j.paid.2020.110350

Ames, D. R., Kammrath, L. K., Suppes, A., & Bolger, N. (2010). Not so fast: The (not-quite-complete) dissociation between accuracy and confidence in thin-slice impressions. *Personality and Social Psychology Bulletin*, *26*(2), 264–277. https://doi.org/10.1177/0146167209354519

Amir, O., Rand, D. G., & Gal, Y. K. (2012). Economic games on the internet: The effect of $1 stakes. *PLoS ONE*, *7*(2), 1–4. https://doi.org/10.1371/journal.pone.0031461

Back, M. D., & Nestler, S. (2016). Accuracy of judging personality. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The Social Psychology of Perceiving Others Accurately* (1st ed., pp. 98–124). Cambridge University Press. https://doi.org/10.1017/CBO9781316181959.005

Banerjee, S., Galizzi, M. M., & Hortala-Vallve, R. (2021). Trusting the Trust Game: An External Validity Analysis with a UK Representative Sample. *Games*, *12*(3), 66. https://doi.org/10.3390/g12030066

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Beer, A. (2014). Comparative personality judgments: Replication and extension of robust findings in personality perception using an alternative method. *Journal of Personality Assessment*, *96*(6), 610–618. https://doi.org/10.1080/00223891.2013.870571

Beer, A., & Watson, D. (2010). The effects of information and exposure on self-other agreement. *Journal of Research in Personality*, *44*(1), 38–45. https://doi.org/10.1016/j.jrp.2009.10.002

Berry, D. S., & Zebrowitz-McArthur, L. A. (1988). What's in a face? Facial maturity and the attribution of legal responsibility. *Personality and Social Psychology Bulletin, 14*(1), 23–33.

Biesanz, J. C. (2010). The social accuracy model of interpersonal perception: Assessing individual differences in perceptive and expressive accuracy. *Multivariate Behavioral Research*, *45*, 853–885. https://doi.org/10.1080/00273171.2010.519262

Biesanz, J. C., & Human, L. J. (2010). The cost of forming more accurate impressions: Accuracy-motivated perceivers see the personality of others more distinctively but less normatively than perceivers without an explicit goal. *Psychological Science*, *21*(4), 589–594. https://doi.org/10.1177/0956797610364121

Biesanz, J. C., Human, L. J., Paquin, A.-C., Chan, M., Parisotto, K. L., Sarracino, J., & Gillis, R. L. (2011). Do we know when our impressions of others are valid? Evidence for realistic accuracy awareness in first impressions of personality. *Social Psychological and Personality Science*, *2*(5), 452–459. https://doi.org/10.1177/1948550610397211

Bird, B. M., Moreau, B. J. P., Arnocky, S., & Carre, J. M. (2017). The facial width-to-height ratio predicts sex drive, sociosexuality, and intended infidelity. *Archives of Sexual Behavior*. https://doi.org/10.1007/s10508-017-1070-x

Bjornsdottir, R. T., Rule, N. O., & Publication, O. F. (2017). The visibility of social class from facial cues. *Journal of Personality and Social Psychology*, *113*(4), 530–546.

Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2015). Face-ism and kernels of truth in facial inferences. *Trends in Cognitive Sciences*, *19*(8), 421–422. https://doi.org/10.1016/j.tics.2015.05.002

Borkenau, P., Brecke, S., Möttig, C., & Paelecke, M. (2009). Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality*, *43*(4), 703–706. https://doi.org/10.1016/j.jrp.2009.03.007

Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology*, *62*(4), 645–657.

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: From behavior to neural mechanism. *Annual Review of Psychology*, *66*, 83–113. https://doi.org/10.1146/annurev-psych-010814-015044

Bovet, J., Tognetti, A., & Pollet, T. V. (2022). Methodological issues when using face prototypes: A case study on the Faceaurus dataset. *Evolutionary Human Sciences*, *4*, e48.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Univer. California Press.

Burris, C., & Edwards, S. (2017). Does facial width-to-height ratio predict male offender aggression? *Journal of Criminal Psychology*, JCP-03-2017-0013. https://doi.org/10.1108/JCP-03-2017-0013

Capozzi, F., Human, L. J., & Ristic, J. (2020). Attention promotes accurate impression formation. *Journal of Personality*, *88*(3), 544–554. https://doi.org/10.1111/jopy.12509

Carré, J. M., & McCormick, C. M. (2008). In your face: Facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1651), 2651–2656. https://doi.org/10.1098/rspb.2008.0873

Carré, J. M., McCormick, C. M., & Mondloch, C. J. (2009). Facial structure is a reliable cue of aggressive behavior. *Psychological Science*, *20*(10), 1194–1198. https://doi.org/10.1111/j.1467-9280.2009.02423.x

Chan, M., Rogers, K. H., Parisotto, K. L., & Biesanz, J. C. (2011). Forming first impressions: The role of gender and normative accuracy in personality perception. *Journal of Research in Personality*, *45*(1), 117–120. https://doi.org/10.1016/j.jrp.2010.11.001

Chua, K. W., & Freeman, J. B. (2021). Facial stereotype bias is mitigated by training. *Social Psychological and Personality Science*, *12*(7), 1335–1344. https://doi.org/10.1177/1948550620972550

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*(2), 322–331. https://doi.org/10.1037/0022-3514.81.2.322

De Neys, W., Hopfensitz, A., & Bonnefon, J. F. (2017). Split-second trustworthiness detection from faces in an economic game. *Experimental Psychology*, *64*, 231–239. https://doi.org/10.1027/1618-3169/a000367

de Vries, R. E., Barends, A. J., & de Kock, F. S. (2021). Dispositional insight: Its relations with HEXACO personality and cognitive ability. *Personality and Individual Differences*, *173*, 110644. https://doi.org/10.1016/j.paid.2021.110644

Deaner, R. O., Goetz, S. M. M., Shattuck, K., & Schnotala, T. (2012). Body weight, not facial width-to-height ratio, predicts aggression in pro hockey players. *Journal of Research in Personality*, *46*(2), 235–238. https://doi.org/10.1016/j.jrp.2012.01.005

DeBruine, L. M. (2020). *Composite images*. https://debruine.github.io/posts/composite-images/

DeSteno, D., Breazeal, C., Frank, R. H., Pizarro, D., Baumann, J., Dickens, L., & Lee, J. J. (2012). Detecting the trustworthiness of novel partners in economic exchange. *Psychological Science*, *23*(12), 1549–1556. https://doi.org/10.1177/0956797612448793

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

Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. *Review of Financial Studies*, *25*(8), 2455–2483. https://doi.org/10.1093/rfs/hhs071

Durkee, P. K., & Ayers, J. D. (2021). Is facial width-to-height ratio reliably associated with social inferences? *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2021.06.003

Feingold, A. (1992). Good-looking people are not what we think. *Psychological Bulletin*, *111*(2), 304–341.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, *116*(3), 429–456. https://doi.org/10.1037//0033-2909.116.3.429

Foo, Y. Z., Sutherland, C. A. M., Burton, N. S., Nakagawa, S., & Rhodes, G. (2022). Accuracy in facial trustworthiness impressions: Kernel of truth or modern physiognomy? A meta-analysis. *Personality and Social Psychology Bulletin*. https://doi.org/10.1177/01461672211048110

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review*, *21*, 1180–1187. https://doi.org/10.3758/s13423-014-0601-x

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, *21*(3), 177–182. https://doi.org/10.1177/0963721412445309

Galizzi, M. M., & Navarro-Martinez, D. (2019). On the External Validity of Social Preference Games: A Systematic Lab-Field Study. *Management Science*, *65*(3), 976–1002. https://doi.org/10.1287/mnsc.2017.2908

Gillath, O., Bahns, A. J., Ge, F., & Crandall, C. S. (2012). Shoes as a source of first impressions. *Journal of Research in Personality*, *46*(4), 423–430. https://doi.org/10.1016/j.jrp.2012.04.003

Gomulya, D., Wong, E. M., Ormiston, M. E., & Boeker, W. (2017). The role of facial appearance on CEO selection after firm misconduct. *Journal of Applied Psychology*, *102*(4), 617–635. http://dx.doi.org/10.1037/apl0000172

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Graham, J. R., Harvey, C. R., & Puri, M. (2016). A corporate beauty contest. *Management Science*, *63*(9), 3044–3056. https://doi.org/10.1287/mnsc.2016.2484

Graziano, W. G., & Tobin, R. M. (2002). Agreeableness: Dimension of personality or social desirability artifact? *Journal of Personality*, *70*(5), 695–728. https://doi.org/10.1111/1467-6494.05021

Green, P., & Macleod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*, 493–498. https://doi.org/10.1111/2041-210X.12504

Hall, J. A., Back, M. D., Nestler, S., Frauendorfer, D., Schmid Mast, M., & Ruben, M. A. (2018). How do different ways of measuring individual differences in zero-acquaintance personality judgment accuracy correlate with each other? *Journal of Personality*, *86*(2), 220–232. https://doi.org/10.1111/jopy.12307

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

Hehman, E., Xie, S. Y., Ofosu, E. K., & Nespoli, G. A. (2018). *Assessing the point at which averages are stable: A tool illustrated in the context of person perception*. https://psyarxiv.com/2n6jq/

Human, L. J., & Biesanz, J. C. (2013). Targeting the Good Target: An Integrative Review of the Characteristics and Consequences of Being Accurately Perceived. *Personality and Social Psychology Review*, *17*(3), 248–272. https://doi.org/10.1177/1088868313495593

Human, L. J., Biesanz, J. C., Finseth, S. M., Pierce, B., & Le, M. (2014). To thine own self be true: Psychological adjustment promotes judgeability via personality-behavior congruence. *Journal of Personality and Social Psychology*, *106*(2), 286–303. https://doi.org/10.1037/a0034860

Human, L. J., Rogers, K. H., & Biesanz, J. C. (2021). In person, online, and up close: The cross-contextual consistency of expressive accuracy. *European Journal of Personality*, *35*(1), 120–148. https://doi.org/10.1002/per.2272

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

Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2022). Understanding the role of faces in person perception: Increased reliance on facial appearance when judging sociability. *Journal of Experimental Social Psychology*, *100*, 104288.

Jaeger, B., & Jones, A. L. (2021). Which facial features are central in impression formation? *Social Psychological and Personality Science*. https://doi.org/10.1177/19485506211034979

Jaeger, B., Oud, B., Williams, T., Krumhuber, E. G., Fehr, E., & Engelmann, J. B. (2022). Can people detect the trustworthiness of strangers based on their facial appearance? *Evolution and Human Behavior*, *43*(4), 296–303.

Jaeger, B., Sleegers, W. W. A., Evans, A. M., Stel, M., & van Beest, I. (2019). The effects of facial attractiveness and trustworthiness in online peer-to-peer markets. *Journal of Economic Psychology*, *75*. https://doi.org/10.1016/j.joep.2018.11.004

Jaeger, B., Todorov, A. T., Evans, A. M., & van Beest, I. (2020). Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions. *Journal of Experimental Social Psychology*, *90*, 104004. https://doi.org/10.1016/j.jesp.2020.104004

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy. In *Handbook of Personality: Theory and Research* (pp. 114–158). https://doi.org/10.1016/S0191-8869(97)81000-8

Jones, A. L., & Jaeger, B. (2019). Biological bases of beauty revisited: The effect of symmetry, averageness, and sexual dimorphism on female facial attractiveness. *Symmetry*, *11*(2). https://doi.org/10.3390/SYM11020279

Jones, A. L., Kramer, R. S. S., & Ward, R. (2012). Signals of personality and health: The contributions of facial shape, skin texture, and viewing angle. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(6), 1353–1361. https://doi.org/10.1037/a0027078

Jones, B. C., DeBruine, L. M., Flake, J. K., Aczel, B., Adamkovic, M., Alaei, R., Alper, S., Álvarez Solas, S., Andreychik, M. R., Ansari, D., Arnal, J. D., Babincák, P., Balas, B., Baník, G., Barzykowski, K., Baskin, E., Batres, C., Beaudry, J. L., Blake, K. R., … Chartier, C. R. (2021). To which world regions does the valence-dominance model of social perception apply? *Nature Human Behaviour*. https://psyarxiv.com/n26dy/

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Jünger, J., Kordsmeyer, T. L., Gerlach, T. M., & Penke, L. (2018). Fertile women evaluate male bodies as more attractive, regardless of masculinity. *Evolution and Human Behavior*, *39*(4), 412–423. https://doi.org/10.1016/j.evolhumbehav.2018.03.007

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. Guilford Press.

Kenny, D. A., & West, T. V. (2008). Zero acquaintance: Definitions, statistical model, findings, and process. In N. Ambady & J. J. Skowronski (Eds.), *First impressions* (pp. 129–146). Guilford Press.

Kerr, L. G., Borenstein-Laurie, J., & Human, L. J. (2020). Are some first dates easier to read than others? The role of target well-being in distinctively accurate first impressions. *Journal of Research in Personality*, *88*. https://doi.org/10.1016/j.jrp.2020.104017

Kordsmeyer, T. L., Lohöfener, M., & Penke, L. (2018). Male facial attractiveness, dominance, and health and the interaction between cortisol and testosterone. *Adaptive Human Behavior and Physiology*, *5*(1), 1–12.

Kosinski, M. (2017). Facial width does not predict self-reported behavioral tendencies. *Psychological Science*, *28*(11), 1675–1682. https://doi.org/10.1177/0956797617716929

Kramer, R. S. S., & Ward, R. (2010). Internal facial features are signals of personality and health. *Quarterly Journal of Experimental Psychology*, *63*(11), 2273–2287. https://doi.org/10.1080/17470211003770912

Krzyzaniak, S. L., Colman, D. E., Letzring, T. D., McDonald, J. S., & Biesanz, J. C. (2019). The effect of information quantity on distinctive accuracy and normativity of personality trait judgments. *European Journal of Personality*, *33*(2), 197–213. https://doi.org/10.1002/per.2196

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). *lmerTest: Tests in linear mixed effects models* [Computer software].

Lang, F. R., Lüdtke, O., & Asendorpf, J. B. (2001). Validity and psychometric equivalence of the German version of the Big Five Inventory in young, middle-aged and old adults. *Diagnostica*, *47*(3), 111–121. https://doi.org/10.1026//0012-1924.47.3.111

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hauam, M., Smoot, M., Bigler, R., Buss, D., Cohen, D., Feingold, A., Holden, G., Kalick, D., Miller, P., & Swann, W. B. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin*, *126*(3), 390–423. https://doi.org/10.1037//0033-2909.126.3.390

Lebreton, M., Langdon, S., Slieker, M. J., Nooitgedacht, J. S., Goudriaan, A. E., Denys, D., Holst, R. J. V., & Luigjes, J. (2018). Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. *Science Advances*, *4*, eaaq0668.

Letzring, T. D. (2008). The good judge of personality: Characteristics, behaviors, and observer accuracy. *Journal of Research in Personality*, *42*(4), 914–932. https://doi.org/10.1016/j.jrp.2007.12.003

Letzring, T. D., Murphy, N. A., Allik, J., Beer, A., Zimmermann, J., & Leising, D. (2021). The judgment of personality: An overview of current empirical research findings. *Personality Science*, *2*, e6043. https://doi.org/10.5964/ps.6043

Li, Q., Heyman, G. D., Mei, J., & Lee, K. (2017). Judging a book by its cover: Children's facial trustworthiness as judged by strangers predicts their real-world trustworthiness and peer relationships. *Child Development*, 1–14. https://doi.org/10.1111/cdev.12907

Lin, C., Adolphs, R., & Alvarez, R. M. (2018). Inferring whether officials are corruptible from looking at their faces. *Psychological Science*, *29*(11), 1807–1823. https://doi.org/10.1177/0956797618788882

Little, A. C., & Perrett, D. I. (2007). Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology*, *98*(1), 111–126. https://doi.org/10.1348/000712606X109648

Mattarozzi, K., Todorov, A., Marzocchi, M., Vicari, A., & Russo, P. M. (2015). Effects of gender and personality on first impression. *PLoS ONE*, *10*(9), 1–13. https://doi.org/10.1371/journal.pone.0135529

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for common designs. R package version 0.9.12-4.1.* [Computer software].

Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin*, *35*(12), 1661–1671. https://doi.org/10.1177/0146167209346309

Nestler, S., & Back, M. D. (2013). Applications and extensions of the Lens Model to understand interpersonal judgments at zero acquaintance. *Current Directions in Psychological Science*, *22*, 374–379. https://doi.org/10.1177/0963721413486148

Nestler, S., Egloff, B., Küfner, A. C. P., & Back, M. D. (2012). An integrative lens model approach to bias and accuracy in human inferences: Hindsight effects and knowledge updating in personality judgments. *Journal of Personality and Social Psychology*, *103*(4), 689–717. https://doi.org/10.1037/a0029461

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, *18*(11), 566–570. https://doi.org/10.1016/j.tics.2014.09.007

Olivola, C. Y., & Todorov, A. (2010a). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, *34*(2), 83–110. https://doi.org/10.1007/s10919-009-0082-1

Olivola, C. Y., & Todorov, A. (2010b). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology*, *46*(2), 315–324. https://doi.org/10.1016/j.jesp.2009.12.002

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

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188. https://doi.org/10.1177/0963721414531598

Paulhus, D. L., Bruce, M. N., & Trapnell, P. D. (1995). Effects of self-presentation strategies on personality profiles and theor structure. *Personality and Social Psychology Bulletin*, *21*(2), 100–108.

Penton-Voak, I. S., Pound, N., Little, A. C., & Perrett, D. I. (2006). Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social Cognition*, *24*(5), 607–640. https://doi.org/10.1521/soco.2006.24.5.607

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

Qiu, L., Lu, J., Yang, S., Qu, W., & Zhu, T. (2015). What does your selfie say about you? *Computers in Human Behavior*, *52*, 443–449. https://doi.org/10.1016/j.chb.2015.06.032

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

Realo, A., Allik, J., Nõlvak, A., Valk, R., Ruus, T., Schmidt, M., & Eilola, T. (2003). Mind-reading ability: Beliefs and performance. *Journal of Research in Personality*, *37*(5), 420–445. https://doi.org/10.1016/S0092-6566(03)00021-7

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLoS ONE*, *7*(3), e34293. https://doi.org/10.1371/journal.pone.0034293

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science, 2*(4), 313–345. https://doi.org/10.1136/bmj.2.3584.509

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology, 104*(3), 409–426. https://doi.org/10.1037/a0031050

Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion, 9*(2), 260–264. https://doi.org/10.1037/a0014681

Satchell, L. P., Davis, J. P., Julle-Danière, E., Tupper, N., & Marshman, P. (2019). Recognising faces but not traits: Accurate personality judgment from faces is unrelated to superior face memory. *Journal of Research in Personality, 79*, 49–58.

Segal, N. L. (2013). Personality similarity in unrelated look-alike pairs: Addressing a twin study challenge. *Personality and Individual Differences*, *54*(1), 23–28. https://doi.org/10.1016/j.paid.2012.07.031

Segal, N. L., Graham, J. L., & Ettinger, U. (2013). Unrelated look-alikes: Replicated study of personality similarity and qualitative findings on social relatedness. *Personality and Individual Differences*, *55*(2), 169–174. https://doi.org/10.1016/j.paid.2013.02.024

Segal, N. L., Hernandez, B. A., Graham, J. L., & Ettinger, U. (2018). Pairs of genetically unrelated look-alikes: Further tests of personality similarity and social affiliation. *Human Nature, 29*, 402–417. https://doi.org/10.1007/s12110-018-9326-2

Shen, X., & Ferguson, M. J. (2021). How resistant are implicit impressions of facial trustworthiness? When new evidence leads to durable updating. *Journal of Experimental Social Psychology, 97*, 104219. https://doi.org/10.1016/j.jesp.2021.104219

Shevlin, M., Walker, S., Davies, M. N. O., Banyard, P., & Lewis, C. A. (2003). Can you judge a book by its cover? Evidence of self-stranger agreement on personality at zero acquaintance. *Personality and Individual Differences*, *35*(6), 1373–1383. https://doi.org/10.1016/S0191-8869(02)00356-2

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Siuda, S., Schlösser, T., & Fetchenhauer, D. (2022). Do we know whom to trust? A review on trustworthiness detection accuracy. *International Review of Social Psychology*, *35*(1). https://doi.org/10.5334/irsp.623

Slepian, M. L., & Ames, D. R. (2015). Internalized impressions: The link between apparent facial trustworthiness and deceptive behavior is mediated by targets' expectations of how they will be judged. *Psychological Science*, *27*(2), 282–288. https://doi.org/10.1177/0956797615594897

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The Life Outcomes of Personality Replication Project. *Psychological Science*, *30*(5), 711–727. https://doi.org/10.1177/0956797619831612

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science*, *21*(3), 349–354. https://doi.org/10.1177/0956797610362647

Stopfer, J. M., Egloff, B., Nestler, S., & Back, M. D. (2014). Personality expression and impression formation in online social networks: An integrative approach to understanding the processes of accuracy , impression management and meta-accuracy. *European Journal of Personality*, *28*, 73–94. https://doi.org/10.1002/per.1935

Sutherland, C. A. M., Liu, X., Zhang, L., Chu, Y., Oldmeadow, J. A., & Young, A. (2017). Facial first impressions across culture: Data-driven modelling of Chinese and British perceivers' unconstrained facial impressions. *Personality and Social Psychology Bulletin*, *44*(4), 521–537. https://doi.org/10.1177/0146167217744194

Sutherland, C. A. M., Rowley, L. E., Amoaku, U. T., Daguzan, E., Kidd-Rossiter, K. A., Maceviciute, U., & Young, A. W. (2015). Personality judgments from everyday images of faces. *Frontiers in Psychology*, *6*, 1–11. https://doi.org/10.3389/fpsyg.2015.01616

Suzuki, A., Tsukamoto, S., & Takahashi, Y. (2017). Faces tell everything in a just and biologically determined world. *Social Psychological and Personality Science*, *10*(1), 62–72. https://doi.org/10.1177/1948550617734616

Todorov, A., Funk, F., & Olivola, C. Y. (2015). Response to Bonnefon et al.: Limited 'kernels of truth' in facial inferences. *Trends in Cognitive Sciences*, *19*(8), 422. https://doi.org/10.1016/j.tics.2015.05.002

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

Van De Ven, N., Bogaert, A., Serlie, A., Brandt, M. J., Denissen, J. J. A., & Serlie, A. (2017). Personality perception based on LinkedIn profiles. *Journal of Managerial Psychology*, *32*(6), 418–429. https://doi.org/10.1108/JMP-07-2016-0220

Van Kleef, G. A. (2010). The emerging view of emotion as social information. *Social and Personality Psychology Compass*, *4*(5), 331–343. https://doi.org/10.1111/j.1751-9004.2010.00262.x

Vazire, S., & Gosling, S. D. (2004). e-Perceptions: Personality impressions based on personal websites. *Journal of Personality and Social Psychology*, *87*(1), 123–132. https://doi.org/10.1037/0022-3514.87.1.123

Verplaetse, J., Vanneste, S., & Braeckman, J. (2007). You can judge a book by its cover: The sequel. A kernel of truth in predictive cheating detection. *Evolution and Human Behavior*, *28*(4), 260–271. https://doi.org/10.1016/j.evolhumbehav.2007.04.006

Vogt, S., Efferson, C., & Fehr, E. (2013). Can we see inside? Predicting strategic behavior given limited information. *Evolution and Human Behavior*, *34*(4), 258–264. https://doi.org/10.1016/j.evolhumbehav.2013.03.003

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wang, D., Nair, K., Kouchaki, M., Zajac, E. J., & Zhao, X. (2019). A case of evolutionary mismatch? Why facial width-to-height ratio may not predict behavioral tendencies. *Psychological Science*, *30*(7), 1074–1081. https://doi.org/10.1177/0956797619849928

Wang, X., & Navarro-Martinez, D. (2023). Increasing the external validity of social preference games by reducing measurement error. *Games and Economic Behavior*, *141*, 261–285. https://doi.org/10.1016/j.geb.2023.06.006

Weisberg, Y. J., De Young, C. G., & Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the Big Five. *Frontiers in Psychology*, *2*(AUG), 1–11. https://doi.org/10.3389/fpsyg.2011.00178

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

Windmann, S., Steinbrück, L., & Stier, P. (2021). Overgeneralizing emotions: Facial width-to-height revisited. *Emotion*. https://doi.org/10.1037/emo0001033

Zebrowitz, L. A. (2012). Ecological and social approaches to face perception. In G. Rhodes, A. Calder, M. Johnson, & J. V. Haxby (Eds.), *Oxford handbook of face perception*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199559053.013.0003

Zebrowitz, L. A. (2017). First impressions from faces. *Current Directions in Psychological Science*, *26*(3), 237–242. https://doi.org/10.1177/0963721416683996

Zebrowitz, L. A., Fellous, J. M., Mignault, A., & Andreoletti, C. (2003). Trait impressions as overgeneralized responses to adaptively significant facial qualities: Evidence from connectionist modeling. *Personality and Social Psychology Review*, *7*(3), 194–215. https://doi.org/10.1207/S15327957PSPR0703_01