The Accuracy and Meta-Accuracy of Personality Impressions from Faces

Bastian Jaeger and Willem W. A. Sleegers
Tilburg University
Julia Stern and Lars Penke
University of Goettingen
Alex L. Jones
Swansea University

Draft version: 01/11/2020
This paper is currently undergoing peer review. Comments are welcome.

Author Note

Bastian Jaeger and Willem W. A. Sleegers, Department of Social Psychology, Tilburg University, The Netherlands; Julia Stern and Lars Penke, Department of Psychology, University of Goettingen, Germany; Alex L. Jones, Department 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: bxjaeger@gmail.com.

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, little is known about people's *meta-accuracy* (i.e., whether they are aware of their judgment accuracy). Even if accuracy is generally low, meta-accuracy would allow people to rely on their impressions in the right situations. In two studies (one preregistered), we examined the accuracy and meta-accuracy of personality impressions. We addressed three crucial limitations of previous studies (a) by incentivizing accuracy and meta-accuracy, (b) by relying on substantially larger samples of raters and targets (646 participants rating 1,660 faces), and (c) by conducting Bayesian analyses to also quantify evidence for the null hypothesis. Our findings consistently suggest that people show neither accuracy nor meta-accuracy when forming face-based personality impressions.

Keywords: social perception; personality impressions; accuracy; meta-accuracy; confidence

The Accuracy and Meta-Accuracy of Personality Impressions from Faces

People form impressions of others' personality based on their facial appearance (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 to make many important decisions, including voting, sentencing, and hiring decisions (Olivola et al., 2014). How problematic is widespread reliance on rapid personality judgments? This question remains a strongly debated topic (Bonnefon et al., 2015; Todorov, Funk, et al., 2015). Whereas some highlight evidence for above-chance accuracy for various traits (De Neys et al., 2017; Lin et al., 2018; Penton-Voak et al., 2006), others point to null findings or argue that accuracy is so low that personality impressions should not be considered a reliable cue (Todorov, Funk, et al., 2015; Vogt et al., 2013). Here, we provide a new perspective. We argue that answering this question not only requires an understanding of how *accurate* people are in inferring personality traits from faces, but also of how *meta-accurate* they are (i.e., whether they are aware of when their judgments are reliable and unreliable). Even if people's impressions are mostly inaccurate, 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, meta-accuracy (sometimes also referred to as accuracy awareness or calibration; Biesanz et al., 2011; Lebreton et al., 2018) can foster adaptive reliance on personality impressions, even if accuracy is relatively low.

The *accuracy question*—whether people's personality judgments from faces correspond to targets' actual personality—has received considerable attention in the literature (e.g., Borkenau & Liebler, 1992; Naumann et al., 2009; Penton-Voak et al., 2006). A host of studies examined correlations between self-reported Big Five personality traits and trait judgments based on facial photographs. However, these studies have yielded inconsistent results. For judgments of extraversion, which usually show the highest levels of accuracy in stranger rating tasks (Kenny & West, 2008), some studies found significant levels of accuracy (Borkenau et al., 2009; Kramer & Ward, 2010; Naumann et al., 2009; Penton-Voak et al., 2006; Satchell et al., 2018), whereas others did not (Ames et al., 2010; A. L. Jones et al., 2012; Shevlin et al., 2003). Similarly inconsistent findings have emerged for judgments of openness, conscientiousness, agreeableness, and emotional stability (Ames et al., 2010; Borkenau et al., 2009; A. L. Jones et

al., 2012; Kramer et al., 2011; Naumann et al., 2009; Penton-Voak et al., 2006; Satchell et al., 2018). Thus, evidence for the accuracy of personality impressions is mixed.

The *meta-accuracy question*—whether people are aware of their impressions being more or less accurate—has received comparatively little attention. Borkenau and colleagues (2009) found that participants 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 by measuring confidence in individual personality impressions (i.e., at the trial level). This allowed them to examine both within-person meta-accuracy (are people more confident when their judgments are more accurate?) and between-person meta-accuracy (are more confident people also more accurate?). They did not find significant evidence for either. These findings converge with studies in which participants formed judgments on the basis of more than a face (e.g., after short interactions or after viewing videos of targets), which suggest that people have limited insights into their judgment accuracy (Biesanz et al., 2011). Overall, despite its theoretical importance, evidence on the meta-accuracy of personality impressions is sparse.

The Current Studies

Here we present the results of two studies (one preregistered) on the accuracy and meta-accuracy of personality impressions from faces. We compare whether participants impressions based on facial photographs are related to self-reported Big Five personality. Participants also indicate how confident they are in the accuracy of their impressions, and we test whether confidence estimates are calibrated. That is, we test whether increased judgment confidence is associated with increased judgment accuracy. We examine both within-person and between-person meta-accuracy.

Our studies address three critical limitations of previous studies on the topic. First, incentives have been shown to improve accuracy and meta-accuracy in a variety of judgments tasks (Botvinick & Braver, 2015; Lebreton et al., 2018). Yet, no studies on personality impressions of the Big Five dimensions incentivized participants' judgments or their judgment confidence. We therefore designed an incentive-compatible judgment task in which participants are incentivized to provide accurate and meta-accurate personality judgments.

Second, the majority of previous findings are based on relatively small samples with 50 or fewer raters or targets. Most relevant for the focus of the current research, the only other study

examining both accuracy and meta-accuracy of personality impressions from faces relied on a sample of 25 raters and 21 targets (Ames et al., 2010). Large samples of raters and targets are crucial for adequate power to detect small effects, but also for testing whether results generalize beyond a specific set of raters and stimuli (Judd et al., 2012). We therefore rely on large 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 judgments in total.

Third, in light of the inconsistent or limited evidence in favor of accuracy and meta-accuracy, it is plausible that personality impressions from faces are neither. Yet, existing studies have exclusively focused on statistical methods that cannot provide evidence for such a null hypothesis. We therefore report the results of Bayesian analyses (alongside frequentist statistics), which can quantify evidence in favor of the null hypothesis (Wagenmakers, 2007).

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.

Study 1

In Study 1, we measured the accuracy and meta-accuracy of Big Five personality judgments. Participants 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 ratings were associated with targets' self-reported trait scores and whether participants were more confident in their ratings when their ratings were actually more accurate. Both accuracy and meta-accuracy 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. The sample size was determined by how many participants completed the study within two weeks. Note that this sample size is considerably larger than the sample sizes of previous studies examining accuracy and meta-accuracy (n = 25, Ames et al., 2010; n = 24 and n = 7, Borkenau et al., 2009). 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).[1]

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). Participants stood in front of a white background and were instructed to display a neutral facial expression. Standing position, lighting, and distance were standardized (for a more detailed description of sample characteristics, see Jünger et al., 2018). Targets' personality was assessed with the 44-item Big Five Inventory (John et al., 2008). Participants 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. Participants were randomly assigned to one image set. We measured participants' personality impressions by asking them to rate the person in the photo on each of the Big Five dimensions. Each dimension was described using two trait adjectives from the Ten-Item Personality Inventory (Gosling et al., 2003). For example, the description for conscientiousness was: "A person who scores low on conscientiousness is disorganized and careless. A person who scores low on conscientiousness is dependable and self-disciplined." Participants provided ratings for one dimension at a time on a scale that ranged from 1 (*not* [trait] *at all*) to 5 (*extremely* [trait]). That is, participants completed 100 trials: Faces were displayed 5 times; each time paired with a different personality dimension. After indicating a trait rating, participants also indicated how confident they were in the accuracy of their rating on a scale that ranged from 1 (*not confident at all*) to 9 (*extremely confident*). 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 (*ICC*s) to estimate consensus in judgments. Across stimulus sets, consensus ranged from *ICC* = .135 to *ICC* = .284 for openness judgments, from *ICC* = .030 to *ICC* = .217 for conscientiousness judgments, from *ICC* = .160 to *ICC* = .248 for extraversion judgments, from

[1] We obtained similar results when including data from excluded participants in our analyses.

ICC = .103 to *ICC* = .222 for agreeableness judgments, from *ICC* = .084 to *ICC* = .267 for emotional stability judgments (all $p < .001$).

Both the accuracy and meta-accuracy of participants' ratings was incentivized independently. Participants were informed that the person with the most accurate ratings (i.e., with the strongest correlation between true and rated personality) and the person with the most meta-accurate ratings (i.e., with the strongest correlation between accuracy and confidence) would each be rewarded with a €50 (ca. $55) voucher for an online retailer.

Analysis strategy. All analyses were conducted in R (R Core Team, 2020). 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). 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 (i.e., a Cauchy distribution with a width of $r = \frac{1}{3}$; 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 proposed by Wagenmakers (2007). We estimated models with and without the variable of interest and computed the Bayesian information criterion (BIC), an indicator of model fit, for both models. Comparing the BICs of both models quantifies the extent to which the variable of interest improved model fit. Following Wagenmakers (2007), we converted this measure to an approximation of the Bayes factor by using 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 variable of interests, 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 $\alpha = 5\%$). Programs commonly used for sensitivity analyses, such as G*Power (Faul et al., 2007), do not support multilevel data. We therefore used the *simr* package (Green & Macleod, 2016) in R (R Core Team, 2020) to conduct sensitivity analyses for the main effects of interest (accuracy across all traits and meta-accuracy

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 model testing accuracy (i.e., the relationship between trait ratings and true scores across all traits) showed that we had 80% power to detect an effect of 0.068. Thus, we could detect a relationship between rated and true personality scores where a one-point increase in true scores is associated with a 0.068-point increase in ratings. Next, we examined power for our model testing meta-accuracy (i.e., the interaction effect between true scores and confidence on trait ratings across all traits). 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 relationship between rated and true personality scores. Thus, our design had sufficient power to detect even low levels of accuracy and meta-accuracy.

Results

Accuracy. First, we examined the accuracy of personality impressions. We estimated a multilevel regression model with random intercepts and slopes per participant and target and regressed participants' trait ratings on the true scores of targets. This did not yield a significant effect and decisive evidence for the null hypothesis, $b = 0.007$, $SE = 0.023$, 95% CI [-0.044, 0.052], $p = .77$, $BF_{01} = 2523$. Across the five personality dimensions, participants did not show accuracy in their impressions. There was significant variation in accuracy across targets, $\chi^2(2) = 155.4$, $p < .001$, but not across participants, $\chi^2(2) = 0.96$, $p = .62$ (see Figure 1). That is, while some targets were judged significantly more accurately than others, we did not find that some judges were significantly more accurate than others.

We also tested for accuracy per personality dimensions. A frequentist analysis indicated that there was significant variation in accuracy across the five dimensions, but a Bayesian analysis indicated decisive evidence in favor of the null hypothesis, $F(4, 7895) = 2.74$, $p = .027$, $BF_{01} = 1.55 \times 10^{18}$. At any rate, associations between trait ratings and true scores were not significant, and there was decisive evidence in favor of the null hypothesis for each of the five dimensions (openness: $b = -0.015$, $SE = 0.065$, 95% CI [-0.171, 0.114], $p = .82$, $BF_{01} = 392.4$; conscientiousness: $b = 0.044$, $SE = 0.048$, 95% CI [-0.059, 0.151], $p = .36$, $BF_{01} = 368.9$; extraversion: $b = -0.019$, $SE = 0.061$, 95% CI [-0.129, 0.104], $p = .75$, $BF_{01} = 420.8$;

agreeableness: b = -0.055, SE = 0.059, 95% CI [-0.162, 0.056], p = .35, BF_{01} = 294.2; emotional stability: b = 0.039, SE = 0.057, 95% CI [-0.093, 0.139], p = .49, BF_{01} = 375.4; see Figure 2). Together these results suggest that participants' impressions were not accurate.

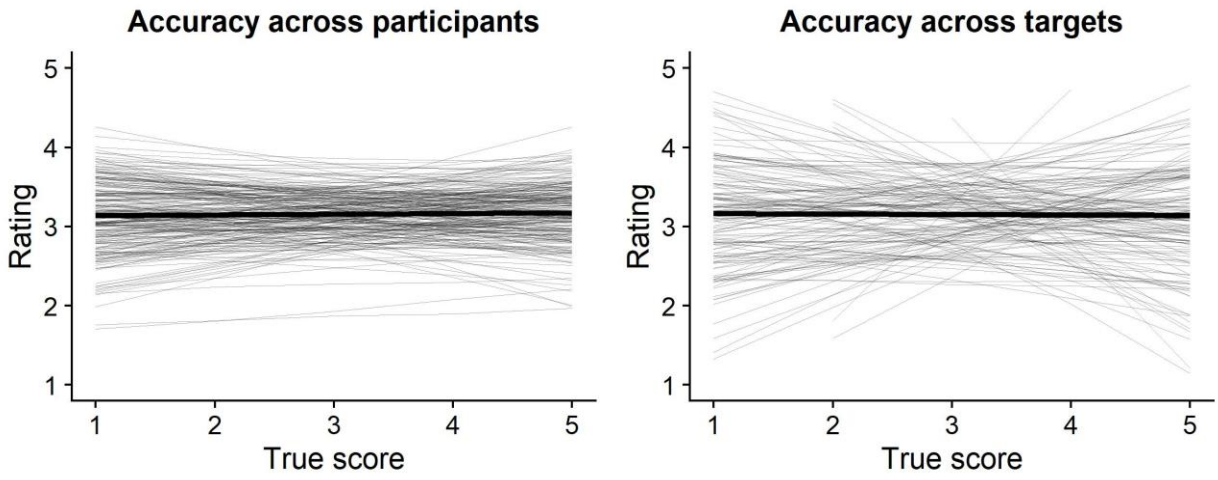

Figure 1. The accuracy of participants' personality impressions. The bold lines in both graphs show the association between true and rated personality traits, visualizing variation in the effect across participants (left) or across targets (right).

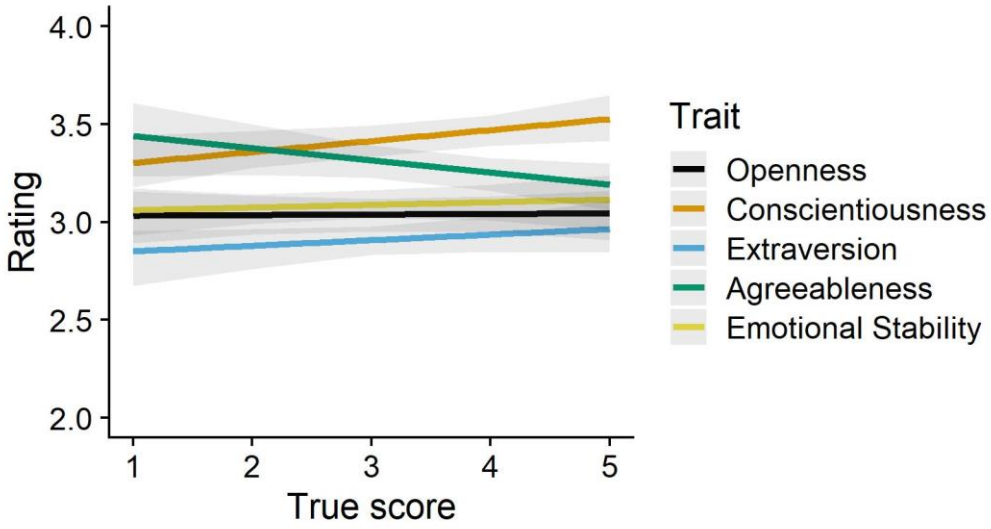

Figure 2. The accuracy of participants' impressions for each personality dimension. The graph displays the associations between participants' trait ratings and targets' true scores.

Meta-accuracy. Next, we examined the meta-accuracy of personality impressions. Participants' mean confidence (averaged across all trials) ranged from 1.10 to 8.99 on our 9-point scale (M = 6.34, SD = 1.65). Participants' 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). Thus, confidence levels varied considerably both between and within participants.

Were participants aware of when their impressions were more or less accurate? We examined this question by testing whether trial-level variation in confidence was associated with trial-level variation in accuracy (i.e., within-person meta-accuracy). In other words, we tested whether there was a stronger association between trait ratings and true scores when participants indicated higher levels of confidence. We estimated a multilevel regression model, in which we predicted trait ratings with true scores, confidence, and their interaction. This did not yield a significant interaction effect, but decisive support for the null hypothesis, b = -0.004, SE = 0.009, 95% CI [-0.022, 0.013], p = .69, BF_{01} = 143172. In other words, accuracy was not higher (i.e., the association between trait ratings and true scores was not stronger) when participants were more confident in the accuracy of their ratings.

We also tested for meta-accuracy per personality dimensions. There was no variation in meta-accuracy across the five personality dimensions with decisive evidence in favor of the null hypothesis, $F(4, 3879) = 1.14$, p = .33, $BF_{01} = 6.83 \times 10^{18}$. Associations between trait ratings and true scores were not moderated by confidence for any of the five dimensions (openness: b = 0.015, SE = 0.018, 95% CI [-0.018, 0.053], p = .40, BF_{01} = 635.8; conscientiousness: b = 0.010, SE = 0.016, 95% CI [-0.024, 0.041], p = .55, BF_{01} = 282.0; extraversion: b = -0.010, SE = 0.019, 95% CI [-0.050, 0.025], p = .58, BF_{01} = 698.7; agreeableness: b = -0.022, SE = 0.017, 95% CI [-0.056, 0.016], p = .21, BF_{01} = 26.69; emotional stability: b = 0.010, SE = 0.017, 95% CI [-0.022, 0.039], p = .54, BF_{01} = 544.8; see Figure 3).

Besides examining the relationship between trial-by-trial variation in accuracy and confidence, we also tested for meta-accuracy at the participant and target level (see Figure 4). There was no significant correlation (with substantial evidence for the null hypothesis) between the average accuracy and confidence of participants (averaged across all targets), $r(221) = -0.02$, p = .78, BF_{01} = 6.17. That is, participants who were more confident were not more accurate. Moreover, there was no significant correlation, and only anecdotal evidence for the null

hypothesis, between the average accuracy and confidence of targets (averaged across all participants), $r(138) = 0.13$, $p = .12$, $BF_{01} = 1.58$. That is, targets that were judged with greater confidence were not judged more accurately. Together, these results suggest that participants were not meta-accurate: Confidence in the accuracy of impressions was not related to the actual accuracy of impressions.

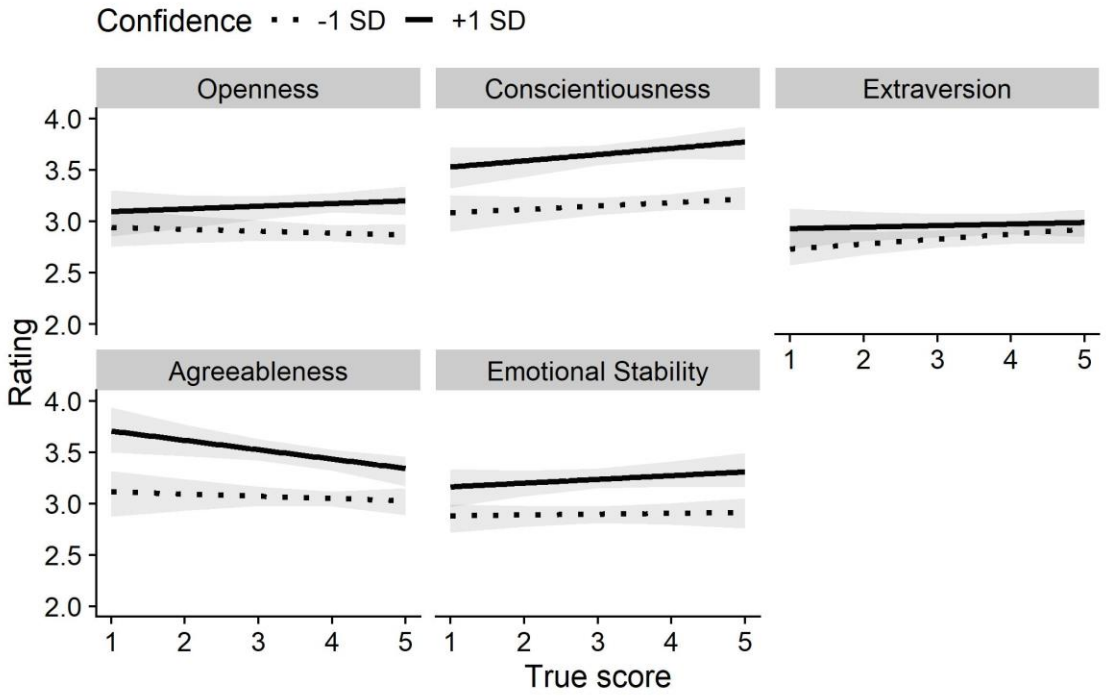

Figure 3. The meta-accuracy of personality impressions. The graph visualizes the relationship between trait ratings and true scores when confidence was low (i.e., one standard deviation below the mean; dotted lines) vs. high (i.e., one standard deviation above the mean; solid lines).

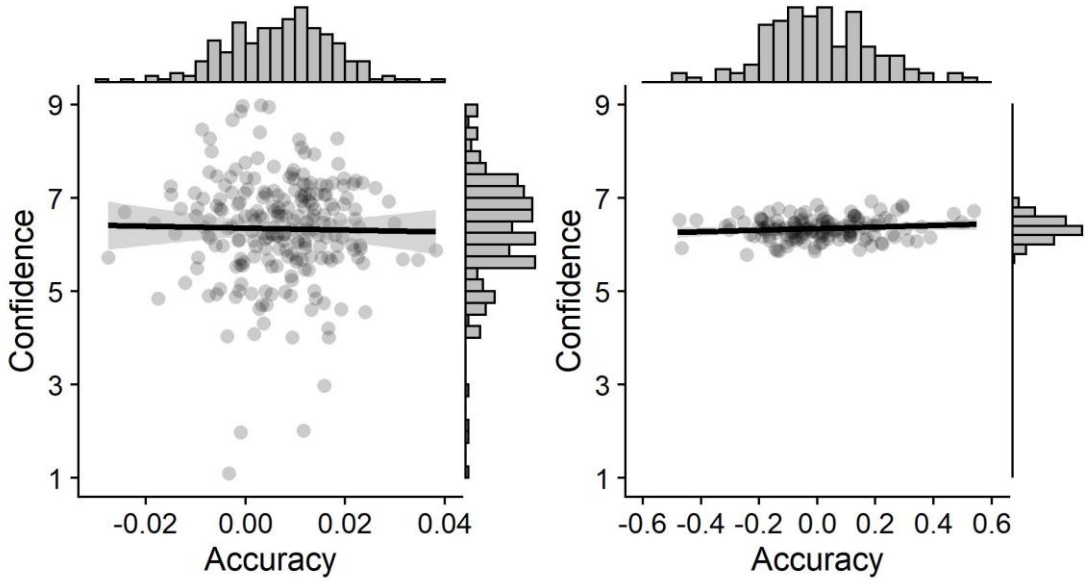

Figure 4. The meta-accuracy of personality impressions at the participant level (left) and at the target level (right). The left graph shows the relationship between the average confidence and accuracy of participants (averaged across all targets). The right graph shows the relationship between the average confidence and accuracy with which targets were judged (averaged across all raters).

Exploratory analyses. We also examined meta-accuracy by analyzing the absolute error in participants' ratings (i.e., the absolute difference between trait ratings and true scores). The average error across all ratings was 1.089 on our 5-point scale (*SD* = 0.792). The average rating error of participants (averaged across all trials) ranged from 0.655 to 1.737 (*SD* = 0.188). There was a positive relationship (with decisive evidence for the alternative hypothesis) between rating error and confidence, $b = 0.054$, $SE = 0.006$, 95% CI [0.040, 0.066], $p < .001$, $BF_{10} = 2.38 \times 10^9$. That is, participants were actually *more* confident when their accuracy was *lower* (i.e., when the discrepancy between trait ratings and true personality scores was larger).

Finally, we explored whether accuracy and meta-accuracy differed between men and women. Predicting trait ratings with true scores, 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], $p = .27$, $BF_{01} = 2871$. In a similar vein, we did not find a significant three-way interaction effect (with decisive evidence for the null hypothesis) between true scores, confidence, and gender, $b = -0.003$, $SE = 0.005$, 95% CI [-0.014, 0.007], $p = .54$, $BF_{01} = 2.30 \times 10^{11}$.

Study 2

We conducted a preregistered replication (see https://osf.io/tr9zp) and extension of Study 1 to test the robustness of our results. We recruited raters from the United States and used an even larger sets of raters (n = 423) and targets (k = 1,260). In Study 2, we focused on extraversion impressions 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 extended the focus of Study 1 in three critical ways. First, we simplified the impression formation task by showing participants pairs of faces and asked them to indicate which person scores higher on extraversion (rather than asking them to indicate a continuous rating). We also adapted our confidence measure by letting participants bet coins on the accuracy of their ratings on each trial. Participants could win coins by betting when their ratings were accurate and we incentivized participants to maximize their total point count. Second, we analyzed judgments of both male and female targets and varied whether participants judged all-male, all-female, or mixed-gender pairs. Third, participants 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

Participants. Simulation results suggest that trait ratings by approximately 20-25 unique raters produce relatively reliable average trait ratings per target (Hehman et al., 2018). We therefore decided to recruit 420 participants, which would result in 30 unique ratings per face pair. Due to the randomization procedure with which participants were matched to face pairs, not all face pairs had 30 unique ratings when we reached our planned sample size. We therefore continued to recruit participants until all face 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 who completed the study in return for $1 and three chances to win a $25 voucher. 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 (for a more detailed description, see Kordsmeyer et al., 2018). Targets' personality was assessed with the German version of the 42-item Big Five Inventory (Lang et al., 2001). Average 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 ($k = 39{,}080$). Face 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 another target from the same sex and 3 times paired with a target from the other sex. Thus, our final stimulus set contained 1,260 face pairs: 420 all-female pairs, 420 all-male pairs, and 420 mixed-gender pairs.

Procedure. Participants completed 90 trials. On each trial, participants saw a randomly drawn face pair. Participants indicated their extraversion impressions by selecting the person that they think is more extraverted. After each rating, we measured participants' confidence. Participants received 10 coins that they could either keep or bet on the accuracy of their rating. When participants bet the coins and their rating was correct, the coins were doubled (i.e., they received 20 coins). When participants bet the coins and their rating was incorrect, the coins were lost (i.e., they received 0 coins). When participants decided not to bet, they received 10 coins. Thus, to maximize their total point count, participants had to bet the coins when they were more confident in the accuracy of their rating and they should keep the coins when they were less confident.

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

We again incentivized the accuracy and meta-accuracy of participants' ratings independently. Participants 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 most meta-accurate ratings (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. We followed the same analysis strategy as in Study 1. For all primary tests, we report the results of frequentist and Bayesian analyses.

Sensitivity analyses. We again used the *simr* package (Green & Macleod, 2016) in R (R Core Team, 2020) to conduct sensitivity analyses for the main effects of interest (accuracy and meta-accuracy of extraversion judgments). Examining power for our model testing accuracy (i.e., the percentage of times participants 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 meta-accuracy (i.e., the relationship 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 meta-accuracy.

Results

Accuracy. We coded participants' ratings as 1 when they were accurate and as 0 when they were inaccurate (i.e., when they did or did not select the more extraverted target). 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 random intercepts per participant and target. 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, it should be noted that a Bayesian analysis indicated strong support in favor of the null hypothesis (i.e., not different from 50%). There was significant variation in accuracy across stimuli, $\chi^2(2) = 3242$, $p < .001$, and across participants, $\chi^2(2) = 4.83$, $p = .028$ (see Figure 5).

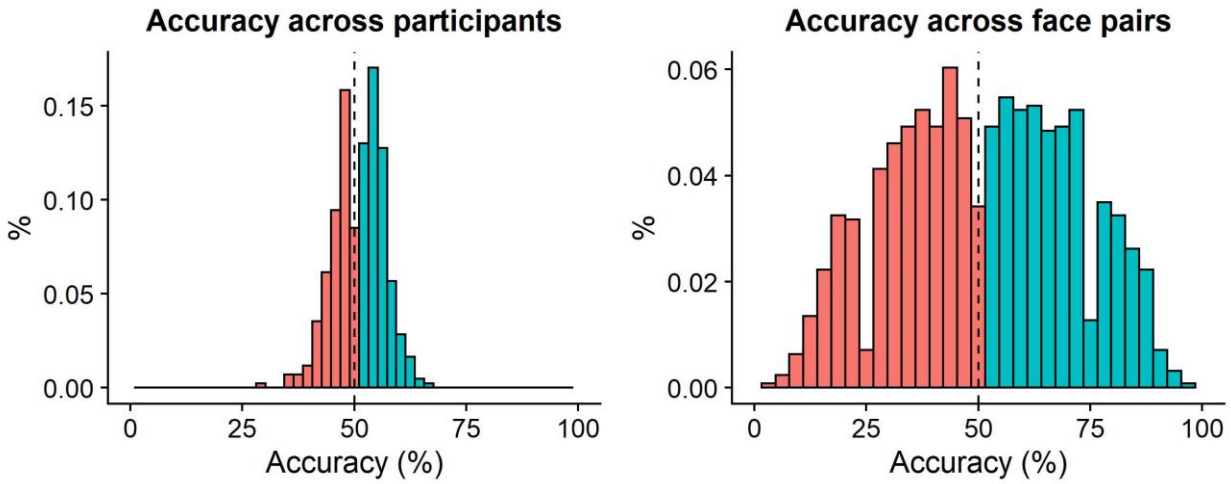

Figure 5. The distribution of accuracy in extraversion impressions across participants (left) and across face pairs (right). Dotted lines denote chance accuracy (i.e., 50%). Participants whose average accuracy across all trials was larger than 50% and face pairs that were judged with more than 50% accuracy (averaged across all raters) are displayed in blue. Participants 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.

Meta-accuracy. Next, we examined participants' confidence in the accuracy of their ratings by analyzing their betting behavior. Participants bet on the accuracy of their rating on 56.00% of all trials, with 41 participants (9.69%) always betting and 22 participants (5.20%) never betting. Participants were incentivized to bet (or not bet) when they thought that their rating was accurate (inaccurate), as this would lead to the highest gains. We realized that in this context, the behavior of participants who always or never bet is difficult to interpret. Both strategies lead to the same earnings if participants believe that their ratings are not accurate at all (50% accuracy). For participants 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 decided to exclude invariant bettors from all analyses of betting decisions, even though this exclusion criterion was not preregistered. However, analyses that included invariant bettors (reported in the Supplemental Materials) led to similar results.

We estimated a multilevel regression model with random intercepts and slopes per participant and target, in which we predicted rating accuracy with betting behavior (0 = did not bet, 1 = did bet). 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 = .77, BF_{01} = 172.3. In

other words, impressions were not more accurate when participants were more confident in them.

We also tested for meta-accuracy at the participant and target level (see Figure 6). The correlation between the betting frequency and accuracy of participants (averaged across all face pairs) was not significant with substantial evidence in favor of the null hypothesis, $r(358) = .03$, $p = .57$, $BF_{01} = 6.91$. That is, participants who were on average more confident were not more accurate. Moreover, the correlation between the betting frequency and accuracy of face pairs (averaged across all participants) was not significant with substantial evidence in favor of the null hypothesis, $r(1258) = .04$, $p = .19$, $BF_{01} = 6.44$. That is, face pairs that were on average judged with greater confidence were not judged more accurately.

Despite this apparent lack of meta-accuracy, participants 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, participants accumulated 912.6 coins ($SD = 70.08$), which was significantly different from 900 (with strong evidence for the null hypothesis), $t(359) = 3.42$, $p < .001$, $d = 0.18$, $BF_{10} = 17.96$.

Finally, we analyzed participants' overall confidence in their performance on the judgment task. On average, participants 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, $r(421) = -.06$, $p = .26$, $BF_{01} = 4.65$, again showing that participants were not meta-accurate (see Figure 7). Moreover, expected accuracy was significantly higher (with decisive evidence in favor of the alternative hypothesis) than actual accuracy, $t(422) = 17.77$, $p < .001$, $d = 0.86$, $BF_{10} = 4.62 \times 10^{29}$. Thus, participants were overconfident in the accuracy of their impressions. More participants overestimated (84.40%), rather than underestimated (14.18%) their accuracy (1.41% provided accurate estimations).

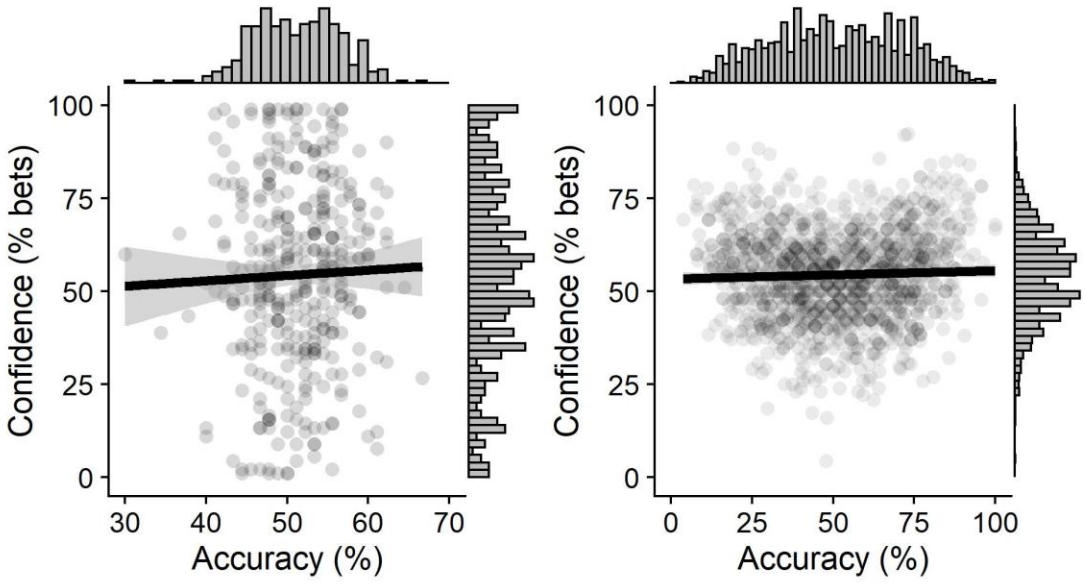

Figure 6. The relationship between betting behavior and accuracy of extraversion impressions at the participant level (left) and at the stimulus level (right). The left graph shows the relationship between the percentage of times participants bet on the accuracy of their ratings and participants' accuracy (averaged across all targets). The right graph shows the relationship between the percentage of times face pairs were bet on and the accuracy with which face pairs were judged (averaged across all raters).

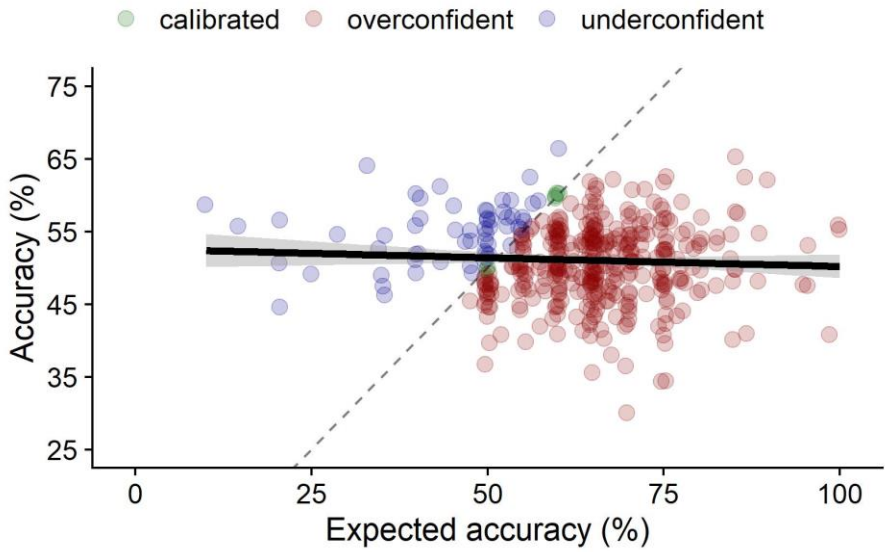

Figure 7. The relationship between the expected and actual accuracy of extraversion impressions. The dashed line represents perfect accuracy (i.e., a correlation between expected and actual accuracy of 1). Deviations to the left of the line signify underestimations of accuracy while deviations to the right signify overestimations over accuracy. Thus, Green dots represent participants that were perfectly meta-accurate, blue dots represent participants that were underconfidence, and red dots represent participants that were overconfident.

Exploratory analyses. We explored whether differences in extraversion scores of targets influenced the accuracy of ratings. If differences in extraversion are reflected in facial features and are therefore to some extent readable by participants, then participants 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 accuracy on extraversion difference 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 = .47, BF_{01} = 149.8. Thus, extraversion impressions were not more accurate when targets actually differed more on extraversion.

Finally, we explored whether accuracy or meta-accuracy varied as a function of target gender (all-male vs. all-female vs. mixed-gender pairs) or participant gender (male vs. female), but found no significant effects and decisive evidence in favor of the null hypotheses (all $BF_{01} \geq$ 5537; see Supplemental Materials for full results).

General Discussion

People form rapid 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 the meta-accuracy of personality impressions—which has received little attention despite its theoretical importance (Ames et al., 2010). Overall, our findings suggest that judges show relatively low levels of consensus when rating personality from faces alone, that personality impressions from faces do not reflect targets' actual personality (i.e., we find no evidence for accuracy), and that people are not aware of when their impressions are more or less accurate (i.e., we find no evidence for meta-accuracy). These conclusions are supported by Bayesian analyses, which yielded (often decisive) evidence in favor of the hypothesis that accuracy and meta-accuracy are not better than chance. Only Study 2 yielded an estimate of 51.10% accuracy for extraversion impressions, which was just significantly higher than chance (i.e., 50%, 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.

Chance-level accuracy and meta-accuracy was obtained (a) for all dimensions of the Big Five, (b) for continuous and binary ratings, (c) irrespective of participant or target gender, and (d) with participants from the Netherlands and the United States. Accuracy and meta-accuracy were not above chance even though we employed judgment tasks that incentivized participants to form accurate and meta-accurate impressions and even though we relied on considerably larger samples of raters and targets compared to previous studies. Moreover, we found no evidence for within-person meta-accuracy (people's judgments were not more accurate when they were more confident in them) or between-person meta-accuracy (people that were on average more confident were on average not more accurate). In fact, comparing participants' estimated with their actual accuracy showed that they were overconfidence in the accuracy of their personality impressions (Study 2).

Is widespread reliance on personality impressions from faces problematic? The current findings suggest an affirmative answer for two reasons. We find that personality impressions from faces are not accurate. 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 meta-accuracy. It should be noted that the current studies focused on impressions based on standardized photographs in which targets displayed a neutral facial expression. Accuracy might be higher for impressions based on richer stimuli, such short face-to-face interactions, videos, or contextualized images such as profile photos (Borkenau & Liebler, 1992; Funder, 2012).

While the current studies provide consistent evidence against accuracy and meta-accuracy, more work on this topic is needed. In both studies, we employed photographs of German targets and examined personality impressions of raters from Western societies. Future studies could test the robustness of our results using more diverse samples or targets and raters (for an example, see B. C. Jones et al., 2020). More work is also needed to explore the accuracy and meta-accuracy of trait judgments. For example, it is still unclear whether trustworthiness impressions from faces are accurate (De Neys et al., 2017; Rule et al., 2013; Vogt et al., 2013) and there is no evidence on meta-accuracy yet. Studies should also employ different types of face stimuli. Cropped images, in which all aspects other than a person's facial features are removed,

ensure that impressions are actually based on facial appearance. However, they lack the ecological validity of more naturalistic photos, which people actually encounter in real life (e.g., profile photos in dating apps, social networking sites, résumés, or case files). In the present study, we balanced these concerns by using facial photos that were standardized (neutral facial expression against a uniform background), but not cropped (targets' hair was still visible). Ultimately, studies using a range of different stimuli are needed to explore variations in impression accuracy.

References

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

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

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

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

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

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

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

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

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

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/

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., 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. (2020). 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., & 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.

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.

Kramer, R. S. S., King, J. E., & Ward, R. (2011). Identifying personality from the static, nonexpressive face in humans and chimpanzees: Evidence of a shared system for signaling personality. *Evolution and Human Behavior*, *32*(3), 179–185. https://doi.org/10.1016/j.evolhumbehav.2010.10.005

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). *lmerTest: Tests in linear mixed effects models* (R package version 2.0-32). https://cran.r-project.org/package=lmerTest

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

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

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

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for common designs. R package version 0.9.12-4.1.* https://cran.r-project.org/package=BayesFactor

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

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

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

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

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

Satchell, L. P., Davis, J. P., Julle-Danière, E., Tupper, N., & Marshman, P. (2018). Recognising faces but not traits: Accurate personality judgment from faces is unrelated to superior face memory. *Journal of Research in Personality*, *79*, 49–58.

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

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

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

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