Image Variability and Face Matching

Taylor Gogan[1], Dr. Jennifer Beaudry[1, 2], & Dr. Julian Oldmeadow[1]

[1]Department of Psychological Sciences; School of Health Sciences, Swinburne University of Technology; Melbourne, Australia

[2]Research Development and Support; Flinders University; Adelaide, Australia

ORCID IDs

Taylor Gogan https://orcid.org/0000-0002-4212-5122

Julian Oldmeadow https://orcid.org/0000-0002-6644-2341

Jennifer Beaudry https://orcid.org/0000-0003-1596-6708

Corresponding author: Taylor Gogan, Department of Psychological Sciences; School of Health Sciences, Swinburne University of Technology, John St, Hawthorn, VIC 3122, Australia. Email: TGOGAN@swin.edu.au

Word count (excluding abstract, references, appendices): 6468

Abstract

This study investigates whether variability in perceived trait judgements disrupts our ability to match unfamiliar faces. In this preregistered study, 174 participants completed a face matching task where they were asked to indicate whether two ambient face images belonged to the same person or different people (17,748 total data points). Participants completed 51 match trials consisting of images of the same person that differed substantially on one trait (either trustworthiness, dominance, or attractiveness) with minimal differences in the alternate traits. Participants also completed 51 mismatch trials which contained two photos of similar-looking individuals. We hypothesised that participants would make more errors on match trials when images differed in terms of attractiveness ratings than when they differed on trustworthiness or dominance. Contrary to expectations, images that differed in terms of attractiveness were matched most accurately, and there was no relationship between the extent of differences in attractiveness ratings and accuracy. There was some evidence that differences in perceived dominance and, to a lesser extent, trustworthiness were associated with lower face matching performance. However, these relationships were not significant when alternate traits were accounted for. The findings of our study suggest that face matching performance is largely robust against variation in trait judgements.

Keywords: Face matching, face perception, first impressions, unfamiliar faces, trait judgements

We are told not to judge a book by its cover, yet we rapidly and automatically make social judgements about others from their faces (Willis & Todorov, 2006). These first impressions can be used to predict election outcomes (Todorov et al., 2005), employment decisions (Gilmore et al., 1986), and dating preferences (Langlois et al., 2000). Although people make countless social judgements, research suggests that dimensions of trustworthiness, dominance, and attractiveness tend to drive first impressions (Sutherland et al., 2013). Beyond forming impressions, we also use people's faces to identify them.

Although people can efficiently identify familiar faces across different settings, they encounter difficulty when tasked with identifying unfamiliar faces (Jenkins et al., 2011). For instance, White et al. (2014) investigated the performance of passport officers on an unfamiliar face matching task (i.e., determining whether two images belonged to the same person or different people). The passport officers were accurate for 71% of match trials (i.e., images of the same person) and 89% of mismatch trials (i.e., images of different people). Surprisingly, the passport officers' performance was comparable to the general population and did not correlate with years of experience. Some researchers suggest that people underestimate the extent to which a face can vary across images (Jenkins et al., 2011), so even small changes in physical appearance can be regarded as evidence for differences in identity (Kramer & Ritchie, 2016).

Various sources of image variance can influence face matching performance. Photos of the same face can vary drastically in appearance from image to image due to changes in such things as expression, lighting, and viewpoint (Burton, 2013). Sources of image variance can make it difficult to perceive two images as belonging to the same face. For instance, even subtle changes in lighting direction (Hill & Bruce, 1996; Longmore et al., 2008) and camera distance (Noyes & Jenkins, 2017) can impair face matching performance. Moreover, people have difficulty matching images of unfamiliar faces from low quality CCTV footage (Burton

et al., 1999; Bruce et al., 2001), which can have important ramifications for forensic investigations. Sources of image variability that disrupt matching unfamiliar faces can also influence first impressions from faces.

Many sources of image variance can influence perceptions of social traits from faces. The first impressions we form from a face can differ from image to image; people can appear trustworthy in one photo yet untrustworthy in another (Gogan et al., 2021). Differences in appearance across photos can influence both perceptions of identity (Jenkins et al., 2011) and perceptions of social traits (Todorov & Porter, 2014). Changes in facial expressions impair perceptions that photos of an unfamiliar face belong to the same identity (Redfern & Benton, 2017) and can inform trait impressions (Oosterhof & Todorov, 2008). For instance, smiling can signal perceptions of trustworthiness (Sutherland et al., 2013) and can increase performance on a face matching task (Mileva & Burton, 2018). Moreover, even slight changes in viewpoint can hinder matching performance for unfamiliar faces (Favelle et al., 2011) and also modulate the intensity of trait impressions (Sutherland et al., 2017). Put together, many of the sources of image variance that influence perceptions of identity can also lead to changes in impressions.

It is not clear whether the processes involved with forming impressions from faces are independent to processing the identity of a face. Bruce and Young's (1986) seminal paper posited that the identity of a face is processed independently to emotional expressions. Facial characteristics that relate to a person's identity are generally stable and consistent across images, whereas facial expressions tend to be more transient and changeable in nature. Haxby et al. (2000) echoed this and proposed that the neural system processes information from stable and changeable cues separately. However, these theories are silent as to where trait impressions lie within these models.

There is reason to think that facial cues relevant for some trait judgements might overlap with cues pertinent to the identity of a face. When we are exposed to different instances of a face, we learn to distinguish image properties that are diagnostic of identity from those that are not diagnostic (Burton et al., 2005; Bruce, 1994). Information that is consistent across different images of a face are generally thought to be more diagnostic of identity than characteristics that change from image to image (Burton et al., 2005). Although many trait judgements can be heavily informed by transient emotional expressions (Zebrowitz & Montepare, 2008), some can be dependent on more stable cues (Hehman et al., 2015). Traits that rely heavily on changeable cues would be expected to vary considerably across images of the same person whereas those dependent on stable cues would be expected to remain relatively consistent across images. Gogan et al. (2021) found that, with ambient photos of faces (i.e., images reflective of the faces we encounter in everyday life), judgements of attractiveness tended to vary less than judgements of trustworthiness and dominance. This suggests that cues to attractiveness judgements might overlap more with stable cues to identity than traits such as dominance and trustworthiness.

To the best of our knowledge, only two studies have explored the influence of trait judgements on face matching performance. Graham and Richie (2019) investigated the influence of different types of glasses (e.g., sunglasses) on both social judgements and face matching performance. Graham and Richie were also interested in whether image pairs with larger differences in perceived social traits would be associated with higher errors rates in match trials. Each of their participants rated faces that were either wearing sunglasses, reading glasses, or no glasses on perceived trustworthiness, competence, and attractiveness. These same participants then completed a face matching task consisting of the same images that they previously rated. Participants were unaware that they would view the images they previously rated again in the matching task and that they would see multiple images of the

same identity. Graham and Richie found no significant correlations between differences in trait ratings between image pairs and accuracy on match trials for any of the social traits. Mileva (2017) conducted a similar—as of now, unpublished—study, which also investigated the relationship between trait judgements and face matching performance. Mileva calculated differences in perceived trustworthiness, dominance, and attractiveness between image pairs, which had been rated in an earlier experiment by an independent set of participants. Similar to Graham and Richie, Mileva found no significant relationships between discrepancies in trait judgments and accuracy on match trials. However, it is worth noting that the pattern of findings aligned with Mileva's predictions—differences in attractiveness ratings were associated with lower accuracy on match trials than were differences in dominance and trustworthiness. The limited research on this issue has not found convincing evidence that differences in trait appearance are related to face matching, and therefore suggest that trait and identity judgments may be independent processes.

Importantly, aspects of their study designs might account for the findings of Mileva (2017) and Graham and Richie (2019). Graham and Richie manipulated the type of glasses worn in the images to test their primary research questions; however, this manipulation may have distorted the relationship between trait differences in image pairs and matching performance. For instance, Graham and Richie found that incongruency in wearing reading glasses or sunglasses impaired matching performance and that wearing sunglasses reduced perceptions of trustworthiness. Additionally, participants viewed the same images in both the trait rating and matching task and were exposed to multiple images of the same identities (due to the glasses manipulation), both of which may have induced familiarity and enhanced matching performance (see Clutterbuck & Johnston, 2005). Further, making social judgements of faces can also facilitate later recognition (Winograd, 1976). Another potential limitation with both Mileva's (2017) and Graham and Richie's (2019) studies was the

inability to single out the effect of differences on one trait without the confounding influence of the other traits, as these were not controlled for. In other words, the relationship between discrepancies in attractiveness judgements and matching performance might have been influenced by differences in other traits.

We aimed to address concerns from past research in our current study. Specifically, to avoid inducing familiarity by asking participants to both rate the images and make matching decisions, we used a subset of ambient images that had been rated by independent participants (Gogan et al., 2021). We also selected image pairs for match trials that differed substantially on one trait with minimal differences in other traits. For example, we selected images that differed in terms of attractiveness with minimal differences in trustworthiness and dominance ratings. It is important to note that we drew upon uncontrolled ambient face images in this study and, as such, the match trials varied on many uncontrolled dimensions that might influence matching performance. However, our interest was not in particular cues or dimensions, but rather whether variations that give rise to different trait impressions are systematically associated with matching performance. Although our image pairs vary on dimensions not central to the traits of interest that might influence matching performance (such as, head tilt or eye gaze), we were interested in whether a relationship between trait impressions and matching performance would emerge despite noise from other sources of image variance.

The aim of the current study was to assess whether face matching performance is more robust against variation in some traits than others. In other words, would it be more difficult to perceive two images as belonging to the same person if they differed in terms of attractiveness than if they differed on trustworthiness? We hypothesised that people would make more errors on match trials when the identities had discrepant trait ratings. Specifically, we predicted that accuracy would be lowest when images differed on attractiveness ratings

and highest when images differed on trustworthiness ratings (with dominance falling in between). We expected differences in attractiveness to be more predictive of accuracy in the matching task than other traits because judgements of attractiveness are thought to be relatively stable across images of a face and therefore might covary with perceptions of identity (Gogan et al., 2021). On the other hand, perceptions of trustworthiness tend to vary substantially across images of a face, and as such are unlikely to be confounded with perceptions of identity (Todorov & Porter, 2014). Importantly, we made no predictions for the mismatch trials (i.e., images of two different people) because we did not consider the trait ratings when selecting these image pairs (please see the Materials section for further explanation).

Methods

Disclosure Statement and Deviations from Preregistration

In this study we report our process for determining sample size, all data exclusions, all manipulations, and all measures (Simmons et al., 2012). The Swinburne University of Technology Human Research Ethics Committee (approval number: 2021431-3700) approved this research and all participants provided informed consent. We preregistered our study (https://osf.io/4e6t5), and all materials, code, and data associated with this study are accessible on the Open Science Framework (OSF; https://osf.io/dm94f/).

We made three deviations from our preregistration. First, we originally planned to use a binary logistic regression; after data collection but prior to data analysis, we opted for a hierarchical linear regression to allow for easier interpretation and comparison to other studies. For completeness, we report the logistic regression analysis in the supplemental materials (S1). Second, in line with our preregistration, we collected data from 182 participants with the goal of obtaining a final sample of 150 participants. However, we had to exclude fewer participants than expected, so our final sample consisted of 174 participants.

Finally, we included the discrepancy scores for the alternate traits as level 2 predictors to control for their influence (see the materials section for further details). We made this decision prior to data collection because although we selected images to have minimal differences in the alternate traits, they could still have some influence, so we opted to further control them statistically. We report the results both with and without these covariates included in the analyses.

Design

This study employed an incomplete 3 (trait discrepancy: trustworthiness, dominance, attractiveness) by 2 (trial type: match or mismatch) within-subjects design. This was an incomplete design because mismatch trials were not crossed with trait discrepancy. Trait discrepancy refers to the social trait that image pairs primarily differed on (for match trials), whereas trial type refers to whether a given image pair depicted the same identity (match trials) or different identities (mismatch trials). The dependent variable for all analyses was face matching accuracy (i.e., percentage of correct responses). All participants completed the same 102 trials, presented in randomised order; for half of these trials, participants were presented with two photos of the same person (i.e., match trials), whereas the remaining trials included two photos of different people (i.e., mismatch trials).

Participants

We used the WebPower R package (Zhang & Yuan, 2018) to determine that a sample size of at least 150 participants was needed to ensure adequate power for a logistic regression. The analysis was set for a two-tailed test with an alpha of 0.05, 80% power, and a normal distribution for the predictors. Please see our preregistration (https://osf.io/4e6t5) and OSF project (https://osf.io/dm94f/files/) for our R code and further explanation of the power analysis.

We recruited 182 participants from a first-year psychology class in a large Australian University who participated in exchange for course credit. In line with our preregistered exclusion criteria, we removed 3 participants who completed less than 75% of the experiment and an additional 5 participants who indicated familiarity with one or more identities. The final sample consisted of 174 participants, most of whom identified as female (138 female; 36 male). The age of the participants ranged from 18 to 58 years (M = 28.20; SD = 9.37).

Materials

Selection of Image Pairs

We selected images from Gogan et al.'s (2021) image database to use in the current study. Gogan et al. had 95 participants rated 340 ambient face images (consisting of 17 individuals) on perceived trustworthiness, dominance, or attractiveness on a scale ranging from 1 to 9, with higher scores denoting higher levels of the trait. The 17 individuals were all Caucasian, in their early- to mid-20s; 9 were female and 8 were male. Each of these individuals is referred to as an identity. Gogan et al. resized the images to 300 X 300 pixels and cropped the images to ensure each face was centred with some visible background (please refer to Figure 1 for an example). Gogan et al.'s image database is available on the OSF (https://osf.io/g3euq/files/). Please refer to Gogan et al. for further information about characteristics of the stimuli.

From Gogan et al.'s (2021) set of images, we compiled image pairs for the match trials (i.e., images belonging to the same identity) by finding photos which differed substantially in terms of ratings on one trait but with minimal differences in the other traits. For instance, we found two images of the same person that differed on attractiveness (i.e., one image with low and the other with high ratings) but had similar ratings of dominance and trustworthiness. We selected images in this way to ensure that any relationship between

differences in one trait and face matching performance are unlikely to be due to differences in another trait.

The first author and a research assistant—who was unaware of our hypotheses—selected images for the match trials. We ranked image pairs in terms of rating discrepancy for each trait and short-listed images with the largest ratios in terms of largest discrepancy in the trait of interest relative to differences in alternate traits. It is worth noting that it was not possible to select image pairs that had no variance in ratings of the alternate traits. Trait dimensions tend to be interrelated to such an extent that even computer modelling will not yield an entirely orthogonal factor structure (Sutherland et al., 2013). Next, we ensured that each image was included only once in the face matching task by removing image pairs that were included in another trial (either match or mismatch). In these instances, we retained the image pair with the largest trait discrepancy.

We also created mismatch trials (i.e., photos of two different people) by pairing images of different identities that were superficially similar in appearance (e.g., same hair style, age, gender, weight etc). Importantly, unlike with the match trials, we did not select images for the mismatch trials based on similarities or differences in trait judgements. Conceptually, there is no reason to expect that images of different faces which are perceived to be similar on a trait dimension would lead to a face matching advantage. Images of two people can be perceived as equally attractive, dominant, or trustworthy and yet look entirely different. We included mismatch trials as distractors, otherwise participants would have quickly realised that all of the trials contained images of the same person; this awareness likely would have influenced their responses.

The face matching task contained 102 image pairs in total: 51 match and 51 mismatch trials. The 51 match trials were comprised of three image sets, with each set containing 17 image pairs which differed substantially on one of the three traits. We used new images of a

given identity across trials so that no image was presented more than once in the face matching task.

Trait Rating Discrepancy Scores

Using trait rating data from Gogan et al. (2021), we calculated trait discrepancy scores for each of the match trials to indicate the extent to which the two images from the same identity differed on a given social trait. To calculate this, for the trait that differed most substantially, we subtracted the rating value of one image from the second image and took the absolute value. As such, larger discrepancy scores denote larger differences in ratings between the two images. We then also conducted this calculation on the alternate traits (traits that the image pairs did not differ on substantially) to include these values in our analyses allowing us to statistically control for these differences.

As shown in Table 1, the descriptive statistic scores were similar for image pairs that were selected based on trustworthiness and attractiveness differences, whereas discrepancy scores for dominance pairs were smaller. The discrepancy in ratings for the alternate traits were similar across conditions.

Table 1

Mean and SD Trait Rating Discrepancies across each Trait of Interest and Alternate Traits

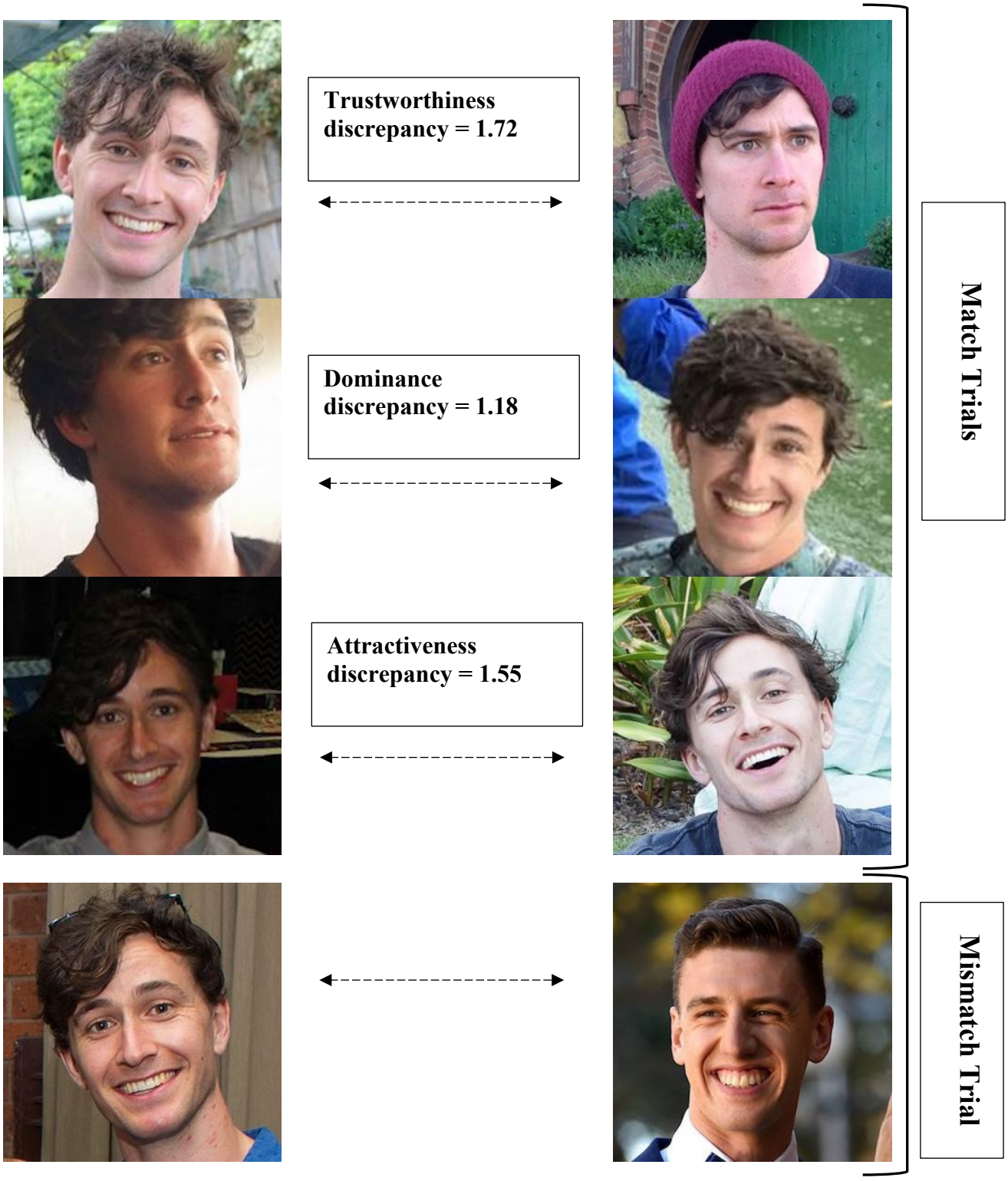
Trait	**Trait rating discrepancy**	
	Mean	**SD**
Trustworthiness	**1.20**	**.31**
Dominance	.23	.19
Attractiveness	.26	.14
Dominance	**.93**	**.29**
Trustworthiness	.24	.18
Attractiveness	.24	.17
Attractiveness	**1.11**	**.36**
Trustworthiness	.17	.13
Dominance	.29	.18

Note. The trait that the image pairs differed most on are bolded whereas the rating discrepancy for the alternate traits are indented. The trait rating discrepancy scores represent the mean and standard deviation of the difference in trait ratings between image pairs.

Figure 1 shows examples of match trials of a pair of images of the same identity that differ substantially on trustworthiness (top row), dominance (second row), and attractiveness judgements (third row). Images with the higher rating of a given pair are displayed on the left and the images with the lower rating are presented on the right. The trait discrepancy scores for each pair are displayed between images of a given pair; the original rating scale ranged from 1 to 9. The bottom row is an example of a mismatch trial.

Figure 1

Example Stimuli of Match Trials for each Trait Condition and a Mismatch Trial



Procedure

We conducted the experiment using Qualtrics (Provo, UT). After providing consent, participants were asked to provide their age, gender, and ethnicity. Participants were then provided the following instructions: "In the following task you will be presented with two face images presented side-by-side. Each of these image pairs will either consist of the same person or photos of two different people. You will be asked to indicate whether the images belong to the same identity or two different identities." Participants then proceeded to complete the face matching task. In each of the 102 trials, two images were presented adjacently in the centre of the screen until participants selected one of the response options: "same person" or "different people". After they entered their response, the next image pair appeared. After completing the matching task, participants were asked to indicate whether they were familiar with any of the individuals that they saw during the experiment. Finally, participants were debriefed and thanked for their time. Participation took 17.3 minutes on average (*SD* = 16.6).

Measures

Correct responses on the face matching task— "same person" responses on match trials and "different people" on mismatch trials—were recorded as 1. Incorrect responses—"different people" responses on match trials and "same person" on mismatch trials—were coded as 0. We aggregated the data such that accuracy was measured as a percentage of correct responses, with higher scores indicating higher accuracy.

Results

Analytic Plan

We processed and analysed the data using R (R Core Team, 2019); please see supplemental materials (S2) for the reproducible code statement. We conducted a hierarchical linear regression analysis to test our hypothesis that participants would be less accurate on

match trials with image pairs that differ on rated attractiveness than on match trials with image pairs that differ on dominance or trustworthiness. Specifically, larger discrepancies in ratings for attractiveness would be associated with lower accuracy in the face matching task. If differences in trait judgements disrupted face matching performance, we would expect larger discrepancy scores to be associated with lower accuracy. We accounted for potential confounding factors such as the actual trait ratings of image pairs and the discrepancy sores for the alternate traits.

Face Matching Performance

On average, participants displayed higher accuracy on mismatch trials (M = 85%, SD = .36) than match trials (M = 80%, SD = .40). We also calculated signal detection measures where hits (n = 7,081) represent "same person" responses to match trials, misses (n = 1,793) were "different people" responses to match trials, false alarms (n =1,366) were "same person" responses to mismatch trials, and correct rejections (n = 7,508) were "different people" responses to mismatch trials. Our d' value[1] of 1.85 showed that our participants had comparable sensitivity to Graham and Richie's (2019) participants who completed the face matching task in the 'no glasses' condition (d' = 1.67). Our participants had a slight bias toward responding "different people" as indicated by the positive criterion value (c = .09) and the beta value (1.19) that exceeded 1 (Stanislaw & Todorov, 1999). In other words, participants were relatively conservative in their responses.

Trait Differences and Face Matching

The following section will focus on the relationship between discrepancy in trait ratings between image pairs and accuracy on match trials. Contrary to our expectations, participants were more accurate when matching image pairs that differed in terms of

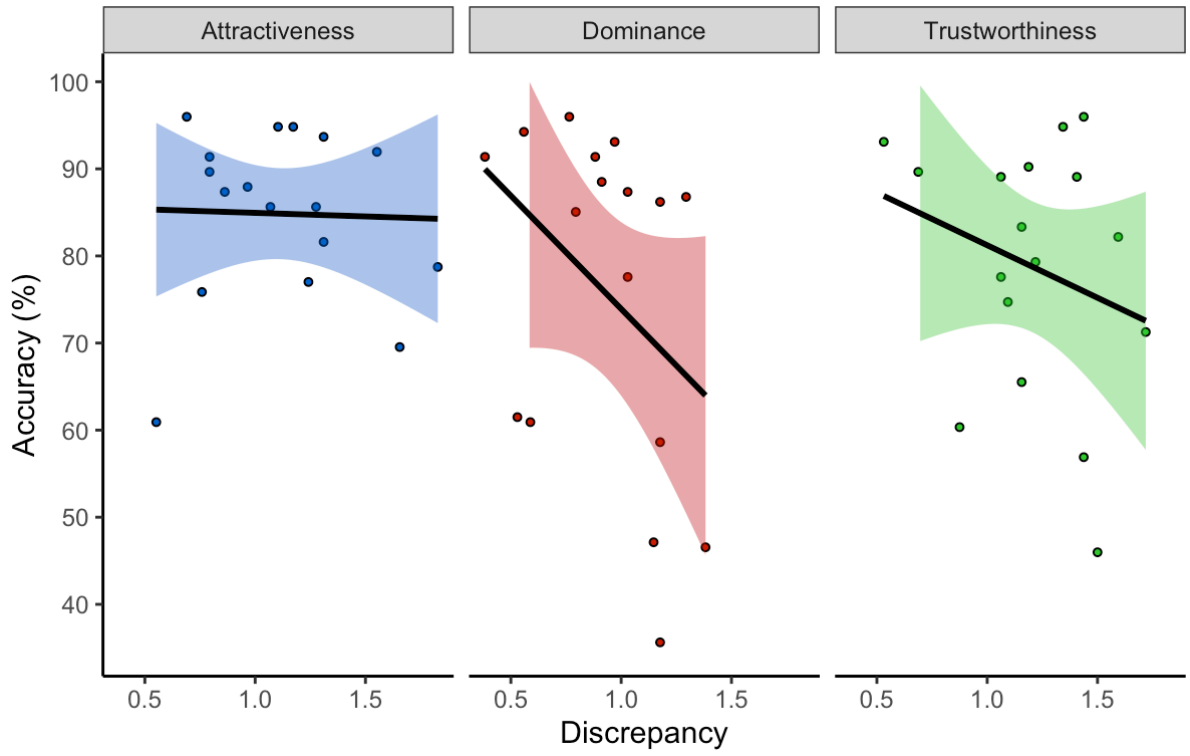
[1] It was not possible to calculate signal detection measures for each of the trait conditions because we manipulated trait differences between image pairs only for the match trials but not mismatch trials.

perceived attractiveness (*M* = 85%; *SD* = .10) than trustworthiness (*M* = 79%; *SD* = .15) or dominance (*M* = 76%; *SD* = .20). Although trustworthiness image pairs had the largest discrepancy scores (see Table 1), participants' face matching accuracy was not the lowest for these images. Additionally, dominance image pairs had the smallest discrepancy scores, yet participants showed lower performance in this condition than the other trait conditions. Taken together, these descriptive statistics do not support our hypothesis that larger differences in trait perceptions between images would lead to lower performance on the matching task.

Figure 2 displays Pearson correlations between accuracy of each match trial and the discrepancy scores for each trait condition. If differences in trait perceptions disrupted face matching performance, we would expect negative relationships between trait discrepancy scores and accuracy. We aggregated the data across participants so that there was a mean accuracy score for each of the 51 match trials. We found no relationship between the discrepancy in attractiveness ratings for image pairs and matching performance $r(15) = -.03$, $p = .909$. There was a negative moderate relationship between discrepancy scores and accuracy for the dominance condition $r(15) = -.39$, $p = .127$, and a negative weak relationship for the trustworthiness condition $r(15) = -.26$, $p = .316$. However, none of these correlations were statistically significant.

Figure 2

Correlations between Rating Discrepancy Scores and Match Trial Accuracy across Each Trait Condition



Note. The shaded bands represent the 95% confidence intervals. Each of the coloured dots represent a matched image pair.

We performed a hierarchical linear regression analysis to assess whether differences in trait ratings between image pairs on match trials would impair performance when controlling for extraneous variables. The regression analysis consisted of three levels. At level 1, we entered the raw trait ratings for both images in each match trial. It was important to control for the actual ratings of the images because there is evidence that faces rated highly on a given trait, such as attractiveness, are better remembered than those with lower ratings (Malloy et al., 2021). In other words, it is possible that images perceived to be high or low on a given trait might influence matching performance irrespective of the difference between the images. At level 2, we included the discrepancy trait rating scores for the traits that the

images did not substantially differ on (i.e., alternate traits) to control for their influence. At level 3, we entered the discrepancy scores for the three traits of interest. The dependent variable was the proportion of correct responses (0–1), with higher scores indicating better matching performance.

Table 2 displays estimates at each level of the hierarchical regression analysis. Level 1 of the regression, which included the raw image ratings was not significant, $F(2, 48) = 0.28$, $p = .759$, and accounted for only 1% of the variance in matching accuracy. The addition of the discrepancy trait rating scores for the alternate traits at level 2 was also not significant, $F(4, 46) = 0.69$, $p = . 606$, and accounted for only an additional 5% of the variance in matching performance beyond level 1. The addition of the level 3 predictors was also not significant, $F(7, 43) = 1.17$, $p = .339$, and accounted for a further 11% of the variance in accuracy scores beyond the previous level. The standardised *Beta* estimates suggest that discrepancy in dominance ratings was the most important predictor in the analysis. The estimates suggest that for every 1 unit increase in rating discrepancy, accuracy decreased by .06 for attractiveness, .18 for dominance, and .12 for trustworthiness pairs.

Given that there were no significant predictors at any stage of the analysis, we ran the analysis again with only the level 3 predictors. The level 3 predictors accounted for 15% of the variance in accuracy overall; however, this model was also not significant, $F(3, 47) = 2.68$, $p = .058$. Of note, dominance discrepancy was the only significant predictor of accuracy; however, this was only the case when the lower-level predictors (specifically level 2) were not controlled for (as in model 3). This suggests that the effect of dominance discrepancy scores on accuracy might, in part, be due to differences in trustworthiness or attractiveness of the images.

Table 2

Parameter Estimates for the Hierarchical Linear Regression

	Level 1		Level 2		Level 3		Level 3 Only	
Predictors	Estimate	Beta	Estimate	Beta	Estimate	Beta	Estimate	Beta
Intercept	0.69 ***	0.00	0.72 ***	0.00	0.88 ***	-0.03	0.93 ***	-0.02
Image 1	0.01	0.05	0.01	0.05	-0.01	0.06		
Image 2	0.01	0.09	0.02	0.10	0.02	0.12		
Alternate Trait 1			-0.19	-0.21	-0.05	-0.16		
Alternate Trait 2			0.01	0.01	0.01	0.04		
Attractiveness					-0.06	0.05	-0.07	0.01
Dominance					-0.18	-0.32	-0.19 *	-0.32
Trustworthiness					-0.12	-0.17	-0.12	-0.24
Observations	51		51		51		51	
R^2 / R^2 adjusted	0.011 / -0.030		0.056 / -0.026		0.160 / 0.023		0.146 / 0.091	

** p<0.05 ** p<0.01 *** p<0.001*

Notes. "Image 1" and "Image 2" refer to the raw ratings for each image in a pair; "Alternate Trait 1" and "Alternate Trait 2" refer to the discrepancy scores for the traits that the images did not substantially differ on; the remaining predictors refer to the discrepancy scores for the traits of interest.

Discussion

The aim of our study was to assess whether differences in trait judgements between images were related to impaired performance on a face matching task. We found no support for our hypothesis that images of the same person that differed in terms of attractiveness judgements would be matched less accurately than images that differed on trustworthiness or dominance judgements. In fact, image pairs that differed substantially on attractiveness judgements displayed the highest accuracy, and there was no relationship between the extent of attractiveness discrepancy and performance on the matching task. However, there was

some—albeit weak—evidence that differences in ratings of dominance were associated with impaired performance. Images that differed in terms of dominance had the lowest accuracy, and larger discrepancies in ratings were associated with lower accuracy; however, this was not significant when alternate traits were controlled for. This suggests that although face matching performance showed the most impairment when the images differed primarily in terms of perceived dominance, this was in part due to differences in the perceived trustworthiness and/or attractiveness of the images.

Our findings suggest that variations in attractiveness judgements are not confounded with perceptions of identity. Past research has demonstrated that judgements of attractiveness are relatively consistent across different images of a face (Gogan et al., 2021; Todorov & Porter, 2014). Similarly, Burton et al. (2005) contended that characteristics that are stable across images of a face are diagnostic of identity. It is therefore interesting that image pairs that differed in terms of attractiveness did not impair matching performance. However, it is worth noting that not all characteristics that are stable across different images provide identity-specific information. For instance, characteristics such as race and gender are consistent across images of a face, yet do not provide individuating information (Kramer et al., 2017). Our findings suggest that although judgements of attractiveness and identity are relatively stable across images, these two judgements do not covary.

Our findings were largely concordant with past research (e.g., Graham & Richie, 2019; Mileva, 2017). In line with these studies, we found little evidence that trait judgements (with the possible exception of dominance) related to performance on a face matching task. We expected that deliberately selecting images that differed primarily on a single trait and controlling for the influence of alternate traits would reveal an underlying relationship between specific traits and matching performance. Despite our efforts to address

methodological limitations of previous research, we found limited evidence of an underlying relationship.

We might have found minimal evidence of a relationship between the discrepancy in trait judgements and matching performance because we had independent participants provide the trait ratings and perform the face matching. We selected image pairs based on ratings from Gogan et al. (2021); it is possible that participants in the current study did not perceive meaningful trait differences between images. Some research suggests there are high levels of agreement in trait judgements between observers (Todorov et al., 2015), while other studies contend that personal preferences play a significant role (Hehman et al., 2017; Kramer et al., 2018). Although personal preferences can explain some variance in impressions, there is still a meaningful amount of agreement across observers. Graham and Richie (2019) bypassed this issue by correlating each participant's trait ratings with their performance on the matching task. However, as discussed earlier, this procedure runs the risk of inducing familiarity through prior exposure to the identities before the matching task. The primary reason we used independent raters was to allow us to select match trials to isolate, to the best of our ability, the effect of differences in a single trait. Drawing upon independent raters allowed us to select match trials in this way without inducing familiarity, which might impact face matching performance. This component of our research design was essential to reduce the confounding effects of other traits—a limitation of previous research.

Another explanation is that simultaneous face matching tasks, which we used in the current study and were used in past studies (e.g., Graham & Richie, 2019), rely heavily on perceptual as opposed to memory processes (Richie et al., 2021). In this task, successfully matching images involves simply looking at differences between the two adjacent images. Alternatively, in sequential face matching tasks, images are presented one at a time rather than simultaneously; this sequential task requires faces to be held in memory in order to

compare two images (Menon et al., 2015). Given the different processes underlying simultaneous and sequential face matching tasks, it is possible that variability in trait impressions might have a larger impact when memory processes are involved (e.g., sequential face matching tasks). Future studies may wish to replicate our experiment using a sequential face matching task to assess whether trait variation can impair performance when memory processes are involved.

Limitations and Future Directions

Some limitations of our study should be highlighted. Differences in accuracy between each trait condition might be due to the use of different image pairs. Selecting image pairs that differed substantially on only one trait with minimal variation in alternate traits meant that different images were used in each trait condition. As such, differences in accuracy across trait conditions might be due to image similarity rather than discrepancy in trait judgements. For instance, image pairs used in the dominance condition may have simply been more difficult to match than images used in the other conditions, irrespective of trait differences. Correlating differences in ratings for various traits with accuracy using the same images (as with Graham & Richie, 2019) would hold image differences constant; however, the relationship between accuracy and differences in one trait would be confounded by differences in other traits.

An additional limitation was the limited range of trait discrepancy scores. The average discrepancy in ratings was less than 1.5 on a 9-point scale for all traits. It is possible that the differences in trait ratings between images was not substantial enough to be associated with disrupted face matching performance. However, it was difficult to select images that differed more than 1.5 points on one trait without corresponding large differences on alternate traits, as judgements from faces tend to be highly correlated (Oosterhof & Todorov, 2008). Future research may wish to replicate our study using computer generated faces to more precisely

manipulate facial characteristics to produce large differences in one trait without creating differences in other traits.

A related issue is that although our use of ambient images enhanced the external validity of the study, this came at the cost of experimental control. Given that the images varied on many dimensions, it is possible that image pairs differed on characteristics that are not weighted heavily in terms of social judgements but might influence matching performance. Additionally, uncontrolled variations in the images likely created noise which reduced the prominence of trait differences described in the previous paragraph. Taken together, it is important for future researchers to replicate our approach using more controlled and standardised images. Specifically, using images with minimal differences in characteristics such as expression, luminance, and camera angle would ensure that image pairs differ only on characteristics relevant to the traits of interest and eliminate potential confounding factors. However, although standardised images would enhance experimental control, a wealth of rich social information would be lost (see Sutherland et al., 2013), potentially further reducing the subtle trait differences between images. Moreover, Burton (2013) argued that "eliminating natural variability may be misleading—at worst leading one to believe that some dimension is important, when it is in fact only important within an artificially constrained set of laboratory stimuli" (p.1482). A possible compromise might be to capitalise on advances in AI-generated face images (e.g., https://generated.photos/faces) which can control for some sources of variance (e.g., hair length, expression, age etc) while retaining the naturalistic and ambient nature of the images. We encourage future research to explore alternative approaches that could shed further light on the relationship between trait impressions and face matching performance.

Finally, our study included only 17 match trials for each trait condition because the image database we used contained images of only 17 individuals (8 male and 9 female) aged

in their early- to mid-20s. We were reluctant to include more than one match trial per identity because participants might have become familiar with the faces. Future researchers should incorporate more trials and expand the scope of this research to explore the role of the race and gender of the stimuli and how this might be modulated by characteristics of the perceivers. Past studies have shown that first impressions can depend on interactions between the gender of the face and of the observer (Mileva et al., 2019; Sutherland et al., 2015). For instance, Mattarozzi et al. (2015) found that women tend to perceive faces as more trustworthy than males, especially when judging female faces. Moreover, the race of the stimuli and cultural background of perceivers can also influence impressions from faces (Jones et al., 2021; Sutherland et al., 2018) and performance on a face matching task (Meissner & Brigham, 2001). Increasing the number of trials and including a more diverse set of faces might allow researchers to investigate the influence of trait variation on face matching performance more thoroughly.

Implications

Limitations notwithstanding, the main theoretical implication of our findings is that trait judgements and face matching abilities seem to be unrelated processes. That is, although a face can appear attractive in one image and unattractive in another, this difference in attractiveness is unrelated to people's ability to see these images as belonging to the same person. Similarly, although there was a small effect of dominance and trustworthiness differences on matching performance, these traits also appear to be unrelated to perceptions of identity. Prominent models of face perception (e.g., Bruce & Young, 1986; Haxby et al., 2000) have remained silent regarding the relationship between trait judgements and identity. The findings of our study suggest that trait impressions from faces might be processed in a similar way to emotional expressions—that is, largely independent to processing identity.

Conclusion

To conclude, we selected image pairs belonging to the same person, which differed substantially on one trait with minimal differences on other traits. Contrary to our expectation, we found no evidence that discrepancies in perceived attractiveness had any influence on face matching performance. Although there was some evidence that larger differences in rated dominance and, to a lesser extent, trustworthiness predicted accuracy on a face matching task, these relationships were not significant when covariates were included. Our findings suggest that face matching abilities are largely robust against variation in traits.

Acknowledgements

This research was supported by the Australian Government Research Training Program Scholarship to the first author. The authors would also like to thank Timothy Cikron-Tighe for his assistance with compiling and selecting the stimuli used in our experiment.

Author contributions

Taylor Gogan: conceptualisation, data curation, formal analysis, investigation, methodology, project admin, resources, visualisation, writing – original draft; Jennifer Beaudry: conceptualisation, methodology, supervising, validation, writing – review & editing; Julian Oldmeadow: conceptualisation, methodology, supervising, writing – review & editing.

Data availability statement

The data that support the findings of this study are openly available on the Open Science Framework at https://osf.io/dm94f/files/

Declaration of conflict of interest:

We have no conflicts of interest to declare.

References

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, *10*(3), 243-248. https://doi.org/10.1111%2F1467-9280.00144

Burton, A. M., Jenkins, R., Hancock, P. J., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology*, *51*(3), 256-284. https://doi.org/10.1016/j.cogpsych.2005.06.003

Burton, A. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1467-1485. http://dx.doi.org/10.1080/17470218.2013.800125

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305-327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Bruce, V. (1994). Stability from variation: The case of face recognition the MD Vernon memorial lecture. *The Quarterly Journal of Experimental Psychology*, *47*(1), 5-28. https://doi.org/10.1080/14640749408401141

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207. https://psycnet.apa.org/doi/10.1037/1076-898X.7.3.207

Clutterbuck, R., & Johnston, R. A. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology*, *17*(1), 97-116*, 31*, 985–994. https://doi.org/10.1068/ p3335.

Gilmore, D. C., Beehr, T. A., & Love, K. G. (1986). Effects of applicant sex, applicant physical attractiveness, type of rater and type of job on interview decisions. *Journal of*

Occupational Psychology, *59*(2), 103-109. https://doi.org/10.1111/j.2044-8325.1986.tb00217.x

Favelle, S. K., Palmisano, S., & Avery, G. (2011). Face viewpoint effects about three axes: The role of configural and featural processing. *Perception*, *40*(7), 761-784. https://doi.org/10.1068%2Fp6878

Gogan, T., Beaudry, J., & Oldmeadow, J. (2021). Within-Person Variability in First Impressions From Faces. *Perception, 50*(7), 595-614 https://doi.org/10.1177%2F03010066211019727

Graham, D. L., & Ritchie, K. L. (2019). Making a spectacle of yourself: The effect of glasses and sunglasses on face perception. *Perception*, *48*(6), 461-470. https://doi.org/10.1177%2F0301006619844680

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223-233. https://doi.org/10.1016/S1364-6613(00)01482-0

Hehman, E., Flake, J. K., & Freeman, J. B. (2015). Static and dynamic facial cues differentially affect the consistency of social evaluations. *Personality and Social Psychology Bulletin*, *41*(8), 1123-1134. https://doi.org/10.1177%2F0146167215591495

Hehman, E., Sutherland, C. A., 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. http://dx.doi.org/10.1037/pspa0000090

Hill, H., & Bruce, V. (1996). The effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(4), 986. https://psycnet.apa.org/doi/10.1037/0096-1523.22.4.986

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313-323. https://doi.org/10.1016/j.cognition.2011.08.001

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., ... & Sirota, M. (2021). To which world regions does the valence–dominance model of social perception apply?. *Nature Human Behaviour*, *5*(1), 159-169. https://doi.org/10.6084/m9.figshare.7611443.v1

Kramer, R. S. S., & Ritchie, K. L. (2016). Disguising superman: How glasses affect unfamiliar face matching. *Applied Cognitive Psychology, 30*, 841–845. https://doi.org/10.1002/acp.3261.

Kramer, R. S., Young, A. W., Day, M. G., & Burton, A. M. (2017). Robust social categorization emerges from learning the identities of very few faces. *Psychological Review*, *124*(2), 115. http://dx.doi.org/10.1037/rev0000048

Kramer, R. S., Mileva, M., & Ritchie, K. L. (2018). Inter-rater agreement in trait judgements from faces. *PloS one*, *13*(8), e0202655. https://doi.org/10.1371/journal.pone.0202655

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin*, 126, 390-423.

Longmore, C. A., Liu, C. H., & Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 77.

Malloy, T. E., DiPietro, C., DeSimone, B., Curley, C., Chau, S., & Silva, C. (2021). Facial attractiveness, social status, and face recognition. *Visual Cognition*, *29*(3), 158-179. https://www.tandfonline.com/author/DiPietro%2C+Carissa

Menon, N., White, D., & Kemp, R. I. (2015). Identity-level representations affect unfamiliar face matching performance in sequential but not simultaneous tasks. *Quarterly Journal of Experimental Psychology*, *68*(9), 1777-1793. https://doi.org/10.1080%2F17470218.2014.990468

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3. https://doi.apa.org/doi/10.1037/1076-8971.7.1.3

Mileva, M. (2017). *Within-Person Variability in Social Evaluation* (Doctoral dissertation, University of York).

Mileva, M., & Burton, A. M. (2018). Smiles in face matching: Idiosyncratic information revealed through a smile improves unfamiliar face matching performance. *British Journal of Psychology*, *109*(4), 799-811. https://doi.org/10.1111/bjop.12318

Mileva, M., Kramer, R. S., & Burton, A. M. (2019). Social evaluation of faces across gender and familiarity. *Perception*, *48*(6), 471-486. https://doi.org/10.1177%2F0301006619848996

Noyes, E., & Jenkins, R. (2017). Camera-to-subject distance affects face configuration and perceived identity. *Cognition*, *165*, 97-104. https://doi.org/10.1016/j.cognition.2017.05.012

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

Redfern, A. S., & Benton, C. P. (2017). Expressive faces confuse identity. *i-Perception*, *8*(5), 2041669517731115. https://doi.org/10.1177%2F2041669517731115

Ritchie, K. L., Kramer, R. S., Mileva, M., Sandford, A., & Burton, A. M. (2021). Multiple-image arrays in face matching tasks with and without memory. *Cognition*, *211*, 104632. https://doi.org/10.1016/j.cognition.2021.104632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21-word solution. *Dialogue: The Official Newsletter of the Society for Personality and Social Psychology, 26*(2), 4-7. https://dx.doi.org/10.2139/ssrn.2160588

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137-149.

Sutherland, C. A., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105-118. https://doi.org/10.1016/j.cognition.2012.12.001

Sutherland, C. A., Young, A. W., Mootz, C. A., & Oldmeadow, J. A. (2015). Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated. *British Journal of Psychology*, *106*(2), 186-208. https://doi.org/10.1111/bjop.12085

Sutherland, C. A., Liu, X., Zhang, L., Chu, Y., Oldmeadow, J. A., & Young, A. W. (2018). Facial first impressions across culture: Data-driven modeling of Chinese and British perceivers' unconstrained facial impressions. *Personality and Social Psychology Bulletin*, *44*(4), 521-537. https://doi.org/10.1177%2F0146167217744194

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, 308, 1623- 1626. https://www.science.org/doi/abs/10.1126/science.1110589

Todorov, A., & Porter, J. M. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological science*, *25*(7), 1404-1417. https://doi.org/10.1177%2F0956797614532474

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

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PloS one*, *9*(8), e103510. https://doi.org/10.1371/journal.pone.0103510

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%2Fj.1467-9280.2006.01750.x

Winograd, E. (1976). Recognition memory for faces following nine different judgments. *Bulletin of the Psychonomic Society*, *8*(6), 419-421.

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass*, *2*(3), 1497-1517. https://doi.org/10.1111/j.1751-9004.2008.00109.x

Zhang, Z., & Yuan, K.-H. (2018). *Practical Statistical Power Analysis Using Webpower and R* (Eds). Granger, IN: ISDSA Press. [https://webpower.psychstat.org]