Knowledge of Identity Reduces Variability in Trait Judgments across Face Images

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

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

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

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): 8702

Abstract

Faces vary from image to image, eliciting different judgments of traits and often different judgments of identity. Knowledge that two face images belong to the same person facilitates the processing of identity information across images, but it is unclear if this also applies to trait judgments. In this preregistered study, participants (N = 100) rated the same 340 face images on perceived trustworthiness, dominance, or attractiveness presented in randomised order and again later presented in sets consisting of the same identity. We also explored the role of implicit person theory beliefs in the variability of social judgements across images. We found that judgements of trustworthiness varied less when images were presented in sets consisting of the same identity than in randomised order and were more consistent for images presented later in a set than those presented earlier. However, knowledge of identity had little effect on perceptions of dominance and attractiveness. Finally, implicit person theory beliefs were not associated with variability in social judgements and did not account for effects of knowledge of identity. Our findings suggest that knowledge of identity and perceptual familiarity stabilises judgements of trustworthiness, but not perceptions of dominance and attractiveness.

Keywords: Face perception, first impressions, trait judgements, within-person variability

Knowledge of Identity Reduces Variability in Trait Judgments across Face Images

We are told not to judge a book by its cover, yet we rapidly and automatically make personality judgements from faces (Willis & Todorov, 2006). Social inferences can have many real-world implications, such as prison sentencing decisions (e.g., Willis & Todorov, 2006), voting (e.g., Todorov et al., 2005), and dating preferences (e.g., Langlois et al., 2000). Although the validity of these personality impressions is questionable (Todorov et al., 2015), judgements are generally consistent across perceivers, implying that they draw upon common facial cues (Sutherland et al., 2013). Yet, the same face can appear drastically different from one image to the next. This can lead to identification errors (Bruce, 1982; Jenkins et al., 2011) as well as changes in trait impressions (e.g., Todorov & Porter, 2014). Trait impression and identity perception research has only recently recognised and studied these within-identity variations (Burton, 2013), largely in isolation of one another.

People differ in attractiveness (Penton-Voak et al., 2000) and some photos of a person are more attractive than others (Jenkins et al., 2011). It is reasonable to extend this to other traits too. People select different photos of themselves for different purposes (e.g., social media profile or a dating site; White et al., 2017), suggesting they have some awareness that different images can signal specific social information. What is less clear, however, is whether within-person variability trumps between-person variability. In other words, can a 'good' photo make you appear as attractive as Brad Pitt or Jennifer Aniston? Research in this area tends to suggest that impressions vary as much or more across images of the same face than between different people; however, this may depend on the type of trait being judged (Jenkins et al., 2011; Sutherland et al., 2017). For instance, perceptions of trustworthiness tend to show substantial within-person variability, whereas attractiveness generally varies more between faces; however, variability in perceptions of dominance is less consistent in the literature (Gogan et al., 2021). The relative variability of different social judgements across

images might depend on the extent to which each trait is derived from changeable versus stable image variations (Hehman et al., 2015).

The term ‘stable facial cues’ is used to refer to cues that yield the same judgment (usually identity) irrespective of changes in the image (such as camera angle or lighting; Hehman et al., 2015). Despite changes in virtually every aspect of an image, we may still be able to determine that it is the same person. This problem can be likened to perceptual constancy, where an object appears unchanged despite changes in size or viewing angle (Walsh & Kulikowski, 1998). Unlike most objects, though, faces contain variable features, such as eyebrows and mouths, in addition to changes caused by image properties. In models of face perception (Bruce & Young, 1986; Haxby et al., 2000), these variable features are cognitively and/or neurologically partitioned from stable cues, the former being recruited for social communication (e.g., emotional state) and the latter for identification. Hence, changes in expression have little impact on identification, at least for familiar faces (Calder et al., 2000). Even for unfamiliar faces, changes in emotional expression do not completely negate identity recognition (Young et al., 1986). Relatively little research has examined how trait judgments vary across images of the same person, so it remains unclear whether trait judgments are linked to identity or are largely independent of identity. Given that even supposedly stable cues can vary across instances due to image properties and/or changes in expression (Kramer, 2016), the question is not so much about whether different traits are linked to stable vs variable cues. Instead, we are interested in to what extent are different trait judgements robust to within-person variability and whether this is associated with the stability of identity.

Gogan et al., 2021 investigated the extent to which judgements of trustworthiness, dominance, and attractiveness varied within and between identities. Participants provided trait ratings of 340 ambient images comprised of 17 non-celebrity identities (20 images per

identity). They found that trustworthiness varied more within identities than dominance, which in turn varied more than attractiveness. Despite differences in judgements across images of the same person, there were still meaningful differences in impressions between different people, suggesting some element of trait impressions is attached to identities.

Do Trait Judgements of Faces Stabilise Across Images?

Necessarily, first impressions research focuses on perception of faces which are unfamiliar to the observer. However, in everyday life we rarely form impressions of randomly presented faces on a computer screen without an awareness of the identity; perceivers usually know which faces belong to which identity. For instance, when you go onto a person's social media page (e.g., Facebook), you would expect that most of the images posted on their page belong to the same identity. A key finding in the facial recognition literature is that people can discount within-identity variability when they are exposed to variable images of a familiar face (Andrew et al., 2016). As such, people can identify familiar faces with high accuracy yet are error-prone when it comes to unfamiliar faces (Jenkins et al., 2011). This is because unfamiliar recognition is largely processed in terms of the image, whereas familiar recognition relies more on abstract mental representations of the identity (Armann et al., 2015). It is unclear whether people can disregard changes in appearance when forming impressions following exposure to images of the same face.

There is some evidence that social judgements are more consistent across images of familiar than unfamiliar faces. Mileva et al. (2019) investigated the extent to which various judgements (trustworthiness, dominance, attractiveness, distinctiveness, and extraversion) varied across images of celebrities who were either familiar or unfamiliar (from foreign countries) to the participants. The authors found that social judgements were more consistent across images of familiar than unfamiliar faces. This suggests that the benefit of familiarity for processing the identity of a face might extend to processing social information from faces,

at least when it comes to photos of celebrities. Mileva et al. reasoned that both perceptual (i.e., physical characteristics of the face) and semantic information (i.e., experience with the person) informed trait judgements. However, it is not clear whether perceptual information alone can attenuate within-identity differences in first impressions.

Some research has demonstrated that perceptions of attractiveness tend to gravitate towards previous judgements of the same face, even without semantic knowledge of the identities. Richie et al. (2017) showed participants 10 images of an identity which were either high or low in perceived attractiveness and were then asked to rate an average-level attractive image of the same person. The authors found that participants who viewed highly attractive images rated the target image as more attractive than participants who saw the low attractiveness images. Similarly, Goller et al. (2018) investigated anchoring effects in trait judgements by presenting images in either ascending or descending order based on attractiveness. Goller et al. found that attractiveness ratings for subsequent images were higher when the images were presented in descending as opposed to ascending order. These studies suggest that ratings of the same face tend to be anchored to past judgments when relying on only perceptual information. However, these studies only assessed perceived attractiveness—a physical trait—which might be more reliant on perceptual information than inferences of a person's character (e.g., dominance and trustworthiness).

In the current study, we assessed whether trait judgements vary less across images of non-celebrity faces when presented in sets consisting of the same identity (grouped rating task) compared to images presented in randomised order (ungrouped rating task). Unlike Goller et al. (2018) and Richie et al. (2017), the photos within a given image set in the current study were presented in randomised order rather than ordered based on previous ratings. Within each grouped image set, we also compared consistency in ratings between the first and second half of the images presented to each participant to assess whether ratings stabilise

following exposure to the first 10 images. As such, we hoped to induce two sources of perceptual familiarity—previous exposure to the faces in the ungrouped rating task and the order of the images presented in the grouped task (i.e., images presented later within sets). However, although we used images of unfamiliar non-celebrities to minimize the semantic information available to participants, some top-down contextual information might play a role. Specifically, in the grouped task, participants were informed of the number of identities within each image set which were also given a fake name; this top-down information is known to influence perceptions of identity (see Andrews et al., 2015; Honig et al., 2021). Nonetheless, differences between experimental conditions would largely be attributed to perceptual processes.

Beliefs about Personality and Social Perceptions

People differ in the extent to which they believe that personality traits are stable as opposed to dynamic (Dweck et al., 1995). It is important to investigate how top–down processes relate to variability in trait impressions across images because a person's beliefs about how traits vary may influence their perceptions of traits from faces. Beliefs that personality traits are stable may constrain a person's judgements across images of the same face whereas people who believe that traits are malleable may perceive traits to vary more freely. If the perceiver is aware that different images belong to the same identity, their beliefs about traits may either accentuate or attenuate perceived variability.

There is a paucity of research into the relationship between beliefs about the stability of traits and variability of impressions formed from faces. However, some research has demonstrated a relationship between these beliefs and social perception. Weisbuch et al. (2016) investigated whether exposure to within-person variability in emotional expression would influence implicit beliefs about the stability of personality (Chiu et al., 1997). Participants viewed photos of seven individuals (consisting of seven photos each). Half of the

participants viewed images of the same person displaying the same emotion in each photo, whereas other participants viewed photos of the same seven identities displaying variable expressions. Weisbuch et al. found that participants were more likely to endorse an incremental than an entity view of personality following exposure to multiple identities displaying substantial within-person variability in emotional expression. However, it is not clear how these beliefs relate to inferences of different social traits. In the current study, we investigated this by examining how variability in judgements for different social traits relate to scores on Chiu et al.'s (1997) implicit person theory scale. Our primary goal here was to control for top–down influences on trait judgments. Even in the absence of semantic information, people may perceive less variability in traits across images of the same person because they believe those traits are stable.

The Current Study

The aim of the current study was to investigate whether being aware of the identity of the images can influence consistency of trait judgements. A secondary aim was to explore the role of beliefs about the stability of personality on the variability of trait evaluations; could this be a mechanism underlying any attenuation of ratings when participants are aware of the identity? We used findings from a pilot study (see S1 in Supplemental Materials) to inform various aspects of the study design in the current experiment.

In the pilot study, we compared trait judgement variability of images when participants rated them in an ungrouped rating task (images presented in random order) compared to a grouped rating task (images presented in sets consisting of the same identity). We found no differences in judgement variability between the grouped and ungrouped rating tasks when using a between-subjects design. However, it is possible that the lack of differences was due to participant differences (see Kramer et al., 2018), especially due to the relatively small sample for each trait condition (*ns* = 34, 27, & 31). As such, in the current

study we opted to employ a within-subjects design whereby the same participants completed both the ungrouped and the grouped rating tasks, with a test delay in between. Following the findings of Gogan et al. (2021), we hypothesised that on average, trustworthiness judgements would vary more than dominance, which would vary more than attractiveness (Hypothesis 1). Additionally, we expected to find less variability in judgements for all three traits in the grouped rating task compared to the ungrouped task (Hypothesis 2).

We also compared the variability in ratings for the first and second half of images presented to each participant within a given image set in the grouped rating task. It is likely that participants' mental representation of a given face would only begin to stabilise after several encounters (see Burton et al., 2011; Jenkins & Burton, 2011). Therefore, we hypothesised that trait ratings for the first 10 images presented for a given identity would tend to vary more than the subsequent 10 images in the grouped rating task (Hypothesis 3).

In our pilot study, we found that beliefs about personality were predictive of variability in trait judgements of faces when measured following the rating task. However, exposure to variable face stimuli might have influenced participant beliefs (see Weisbuch et al., 2016). As such, in the current study we measured beliefs prior to viewing the images. If implicit person theory beliefs are associated with trait rating variability in the grouped condition, this might account for any effects of awareness of the identity on trait judgement variability. We hypothesised that higher scores on the implicit person theory scale will be associated with more variable trait judgements of faces in the grouped rating task only (Hypothesis 4).

Methods

Disclosure Statement

We report how we determined our sample size, all data exclusions, all manipulations, and all measures (Simmons et al., 2012). The Swinburne Human Research Ethics Sub

Committee approved this experiment (reference number: 20215499-8040). We preregistered our study (https://tinyurl.com/2p8n7cue) and posted all materials, data, and code on the Open Science Framework (OSF; https://tinyurl.com/2p9b99z8). Please refer to Appendix A for details about deviations from our preregistration.

Design

This study employed a 3 (trait: trustworthiness, dominance, or attractiveness) X 2 (rating task type: grouped or ungrouped) X 2 (within-identity image presentation order: first and second half of images presented within each set) mixed factorial design. We manipulated trait between subjects to avoid carryover effects between different traits (Rhodes, 2006). The rating task type was a within-subjects factor whereby all participants completed Part 1 (ungrouped rating task) and then Part 2 (grouped rating task) after test delay of 1–7 days. In Part 1, participants were asked to rate all 340 images, presented in randomised order, on one of the three social traits. In Part 2, participants rated these same images on the same social trait; however, the images were presented in image sets consisting of the same identity. That is, images were presented in sets of 20 photos consisting of the same face, and participants were told that all images within a set belonged to the same identity. Importantly, the order of each rating task was not counterbalanced. We presented the ungrouped task first because exposing participants to the grouped identity sets first might have biased their responses on the ungrouped sets (i.e., when all 340 images are presented in randomised order). Finally, in the grouped rating task, the images within each set were pseudo-randomised, whereby half of the participants rated images 1–10 of each identity first and the remaining participants rated images 11–20 first. Additionally, we randomised the presentation order of images within each subset of 10 images for a given identity

Participants

We based our sample size estimates on Jones et al. (2021), who established that to have at least 95% power, we would need between 25 to 30 valid participants in each trait condition (see https://osf.io/x7fus/ for their code and data). We recruited 126 participants through a Research Experience Program (REP; $n = 2$) and via Prolific (https://prolific.co/; $n = 124$). We relied heavily on Prolific for recruitment as data collection was slow for REP. Participants recruited through REP consisted of students enrolled in first-year undergraduate psychology subjects, who were compensated course credits upon completion of each part of the study. Participants recruited via Prolific were compensated £2.50 for part 1 and £3.33 for part 2.

We removed 26 participants from our data based on our preregistered exclusion criteria. Specifically, we excluded 9 (7 in part 1; 2 in part 2) participants who completed less than 75% of the experiment, and 3 who indicated familiarity with the stimuli. We removed a further 9 (3 in part 1; 6 in part 2) participants that gave the same response on more than 75% of trials, and 1 participant who displayed negative test–retest correlations between their ratings in each part. The remaining 4 participants were removed due to only completing part 1 of the experiment. The final sample consisted of 100 participants recruited from REP ($n = 2$) and Prolific ($n = 98$) with ages ranging from 18 to 63 years ($M = 24.2$; $SD = 6.4$). Of the final sample, 49 identified as men, 48 as women, 1 as non-binary, 1 as a transgender woman, and 1 participant wrote “not sure yet”.

Materials

Stimuli

We used Gogan et al.’s (2021) image database. The stimuli consisted of 340 ambient, full-colour photos of 17 non-celebrities; the images varied in numerous characteristics (e.g., gender, hair length, age, expression, setting; see Gogan et al., 2021 for further details). The same stimuli were used in both the grouped and ungrouped rating tasks.

As mentioned in the design section, we manipulated the order of the images of each identity in the grouped rating task. Specifically, we randomly allocated half of the participants to view images 1–10 first and the other half to view images 11–20 first. The image numbers are completely arbitrary, so each subset of 10 images is unlikely to differ in terms of the variability of impressions they elicit.

Implicit Person Theory Scale

Chiu et al's. (1997) implicit person theory scale measures beliefs regarding the extent to which people believe personality attributes, as a whole, are malleable (i.e., incremental theorists) as opposed to fixed (i.e., entity theorists). The eight items were rated on a 6-point scale ranging from 1 (strongly disagree) to 6 (strongly agree), with higher scores reflecting an incremental view of personality and lower scores consistent with an entity theory of personality. Please see Appendix B for the full list of items.

Procedure

Following informed consent, participants completed Part 1 of the study. Participants were invited to answer some demographic questions pertaining to their age, gender, and ethnicity. Participants were then asked to enter the last five digits of their mobile number (to link responses between Parts 1 and 2). Participants then completed the implicit person theory scale (Chiu et al., 1997). Participants were then randomly allocated to one of the three trait conditions (trustworthiness = 34; dominance = 34; attractiveness = 32). Following this, participants received instructions for the ungrouped rating task where they were asked to rate a series of images in terms of perceived trustworthiness, dominance, or attractiveness. Participants rated each image on a scale ranging from 1–9, with higher scores reflecting higher levels of the given trait. Upon completion of the rating task, participants were asked to indicate whether they were familiar with any of the identities prior to completing the study. Finally, participants were thanked and reimbursed for their time.

After a 24-hour delay, participants were invited to complete Part 2 of the study. Once Part 2 was made available to the participants, they had a 7-day window to complete the experiment. The average delay between completing each part was 1.35 days (*SD* = .71). Participants were again asked to enter the last five digits of their mobile number and indicate which trait they rated in Part 1; we cross-checked responses to ensure participants rated the same trait in each task. They completed the implicit person theory scale again. Participants then were provided instructions for the grouped rating task (i.e., that images of the same identity will be presented in separate image sets). Each image set (i.e., identity) was labelled with a fake name. Once participants had completed the rating task, they were debriefed and thanked for their time. On average, participants completed part 1 in 26 minutes and part 2 in 25 minutes.

Measures

Following the approach of Gogan et al. (2021), we converted trait ratings to variability scores to assess the extent to which the images varied within image sets and across traits. These scores were calculated separately for each participant and for each trait. We calculated a mean score for each of the identities by averaging the relevant ratings (i.e., trustworthiness, dominance, or attractiveness) across the 20 images. We then subtracted the rating score of each image from the mean score of the corresponding identity. We used the absolute value of the difference score to reflect the degree, rather than direction, of the discrepancy between the image rating and the mean of the image set. As such, each image received a score from each participant that reflected the degree to which its rating differs from the mean of the image set. If images of a face were given similar trait ratings, then the mean variability score for the identity would be small, whereas if they were given a wide range of scores the mean variability score for the set would be larger.

In other words, for each participant, we calculated the variability score for the i^{th} image of the j^{th} identity set, which is the absolute value of each rating for that image subtracted from the mean rating for that identity.

$$V_{ij} = |x_{ij} - \overline{x}_j|$$

V = Variability score

i = image

j = identity

Results

Analytic Plan

We analysed the data with R (R Core Team, 2019). We performed multilevel modelling using the {lme4} package (Bates et al., 2015) calculated descriptive statistics via the {psych} package (Revelle, 2018), used the {emmeans} package to calculate post hoc tests (Lenth, 2019), and used the {ggplot2} package (Wickham, 2016) for data visualisation. We reported the results of this paper in line with Meteyard and Davies's (2020) proposed recommendations for linear mixed effects models.

We performed linear mixed effects models to assess differences in variability across traits (Hypothesis 1) and between the grouped and ungrouped rating tasks (Hypothesis 2). We performed an additional mixed effects analysis using only the grouped rating task data to assess the influence of image presentation order on trait judgement variability (Hypothesis 3). Finally, we conducted multilevel model analyses to test whether implicit person beliefs are associated with variability in judgements of each trait separately for each rating task (Hypothesis 4).

Trait Judgement Variability across Task Type

We performed Pearson correlations between participant ratings in the ungrouped and grouped rating task to assess consistency in judgements after the test delay. There were strong positive and statistically significant correlations between ratings in each task for

trustworthiness $r(32) = .85$, $p < .001$, dominance $r(32) = .78$, $p < .001$, and attractiveness $r(30) = .88$, $p < .001$. Although this was not a true indication of test-retest reliability given that the conditions of the rating tasks differed, the strong consistency in participant ratings between each task provides some confidence in the reliability of the data.

Moreover, we computed intraclass correlation coefficients (ICCs) separately for each trait and rating task to assess agreement across participants. Following the approach of Lavan et al. (2021), we calculated ICC using a Two-Way-Random model with absolute agreement for each trait and task. We found good agreement for ratings of trustworthiness (*ICC* = .84, 95% CI [.81, .86], $p < .001$), dominance (*ICC* = .86, 95% CI [.83, .88], $p < .001$), and attractiveness (*ICC* = .88, 95% CI [.85, .90], $p < .001$) in the ungrouped rating task. Similarly, we found good agreement for trustworthiness (*ICC* = .86, 95% CI [.83, .88], $p < .001$), dominance (*ICC* = .84, 95% CI [.81, .87], $p < .001$), and attractiveness (*ICC* = .87, 95% CI [.84, .90], $p < .001$) in the grouped rating task. The level of rater agreement demonstrated in our study is comparable to similar past research (e.g., Lavan et al., 2021).

We performed linear mixed effects models to assess differences in variability across the three traits (Hypothesis 1) and between each rating task (Hypothesis 2). Our data had three levels: participants and images were level 1 predictors; the identities of the faces were level 2 predictors; rating task type and trait were level 3 predictors. Participants were nested within traits, and images were nested within identities. We entered trait condition and rating task type as fixed effects with both main effects and an interaction term because we expected these variables to systematically influence the data (Winter, 2013). We entered participants, images, identity, face gender, implicit person belief scores, and test delay as random effects with random intercepts. We included participants as random effects to avoid violating the independence assumption (Winter, 2013). We entered the stimuli variables (images and identities) as random effects to allow our findings to generalise beyond the stimuli used in

this study (Judd et al., 2012; Westfall et al., 2016). We added the delay between tests (i.e., number of days) as a random effect as this might account for some variance when comparing variability scores between Parts 1 and 2. Additionally, face gender was added as a random effect given its importance in social judgements (Mileva et al., 2019). We included implicit person beliefs as a random effect to rule this out as an explanation for differences in variability between the ungrouped and grouped rating task. Finally, we used variability scores as the dependent variable. For the mixed effects model analyses, we added 1 to all variability scores to address non-positive data (i.e., some variability scores had a value of 0); for ease of interpretation, we report the actual variability scores.

Next, we detail the models we built. The dependent variable is presented on the left side of the equation (before the tilde). The fixed effects of each model are shown following the tilde, and the random effects are presented in parentheses. The forward slash between the random effect terms denotes the nesting. The Full model (M4) contained an interaction between trait and task type as well as the main effects of each. The reduced model (M3) included the main effects of trait and task type. Models M1 and M2 contained the single fixed effects of task and trait, respectively. All models included the same set of random effects.

Model Equations:

M4: Variability ~ Trait * Task + (1|Participant) + (1|Identity/Image) + (1|Test Delay) + (1|Beliefs) + (1|Face Gender)

M3: Variability ~ Trait + Task + (1|Participant) + (1|Identity/Image) + (1|Test Delay) + (1|Beliefs) + (1|Face Gender)

M2: Variability ~ Trait + (1|Participant) + (1|Identity/Image) + (1|Test Delay) + (1|Beliefs) + (1|Face Gender)

M1: Variability ~ Task + (1|Participant) + (1|Identity/Image) + (1|Test Delay) + (1|Beliefs) + (1|Face Gender)

Notes. Identity = the identity of the images (1–17); Image = image number within each identity (1–20); Trait = trustworthiness/dominance/attractiveness; Task = the type of rating task (ungrouped/grouped); Test delay = the number of days between completing each task (1–7); Beliefs = implicit person theory scores; Face Gender = the gender of the stimuli (male or female); Variability = the difference between the rating of an image from the mean of the identity.

We performed a likelihood ratio test to establish significant main effects and interactions. Specifically, we compared the full model to a reduced model with the term of interest removed; significant differences in fit between two models can be explained by the term(s) that was dropped from. We deemed models with lower AIC values than the null model and significant *p*-values to represent a better fit to the data; we used AIC values because the BIC penalises more complex models (Wagenmakers & Farrell, 2004). The model fit statistics are presented in Table 1.

Table 1

Model Specifications and Fit Statistics for Task Type and Trait Condition

Model Name	**Model Specification**	**Model Fit**			**Likelihood Ratio Test Statistics**			
		AIC	**BIC**	**df**	**Model Comparison**	**df**	χ^2	***p*-value**
M4	Trait X Task	164674	164793	13	M3	2	60.79	**< .001**
M3	Trait + Task	164731	164831	11	-	-	-	-
M2	Trait	164739	164830	10	-	-	-	-
M1	Task	164734	164816	9	-	-	-	-

Notes. The likelihood ratio test statistics compare the fit of the higher order models to subsequent reduced models. Participants = 100; observations = 68,000; AIC = Akaike

information criterion; BIC = Bayesian information criterion. Bold font indicates significant *p*-values.

As seen in Table 1, the Full model (M4) showed a significantly better fit to the data than the reduced model (M3), suggesting an interaction between trait and task type. In other words, the variability of some trait judgements depended more on the type of rating task than others. Given the significant interaction, the lower order main effects included in the three reduced models are of limited interpretability and, as such, are not discussed. The random effects of image, identity, face gender and test delay all explained a negligible amount of variance (<.05) in the data. Similarly, implicit person beliefs did not explain meaningful variance in the data (*variance* < .01; *SD* = .11), suggesting that participant beliefs did not account for differences in rating variability between tasks. Participant differences explained approximately 11% (*SD* = .33) of the variance in variability scores. However, there was a large amount of residual variance not accounted for by our models (*variance* = .65; *SD* = .81). Figure 1 depicts differences in judgement variability between the grouped and ungrouped rating tasks across each of the three traits.

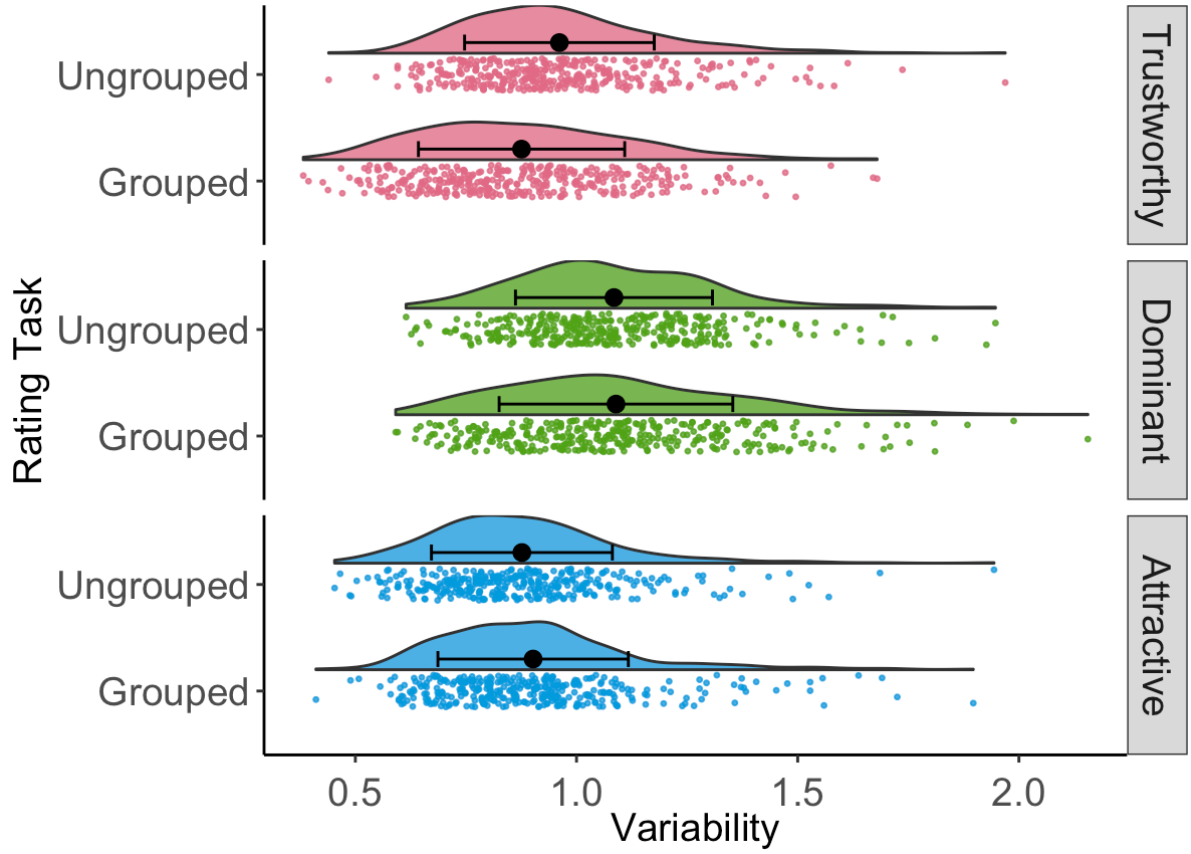
We conducted pairwise post-hoc tests on the full model (M4) to further disentangle the interaction between trait and task type (see Table S2 in Supplemental Materials for descriptive statistics). Degrees of freedom were calculated using the Satterthwaite method. First, we performed post-hoc tests between each of the three trait conditions within each of the rating tasks. We adjusted *p*-values to account for multiple comparisons using Tukey. In the ungrouped rating task, dominance judgements varied significantly more than attractiveness, $t(90.7) = 2.62$, $p = .028$, but there were no significant differences between variability in dominance and trustworthiness judgements, $t(90) = 1.57$, $p = .262$, or trustworthiness and attractiveness judgements, $t(93.7) = 1.08$, $p = .531$. Another pattern of findings emerged in the grouped rating task. Namely, dominance varied significantly more

than trustworthiness, $t(90) = 2.66$, $p = .025$, whereas neither variability in trustworthiness and attractiveness, $t(93.7) = .24$, $p = .970$, nor dominance and attractiveness were significantly different, $t(90.7) = 2.37$, $p = .051$. Put together, dominance varied significantly more than attractiveness in the ungrouped task and significantly more than trustworthiness in the grouped rating task; all other comparisons did not differ significantly.

We also conducted post-hoc tests between each rating task for a given trait. Trustworthiness ratings varied significantly more in the ungrouped than grouped rating task, $t(67558) = 8.07$, $p < .001$. The variability of dominance did not significantly differ by rating task, $t(67558) = 0.43$, $p = .670$. Although attractiveness varied significantly more in the grouped compared to the ungrouped task, $t(67558) = 2.32$, $p = .020$, the mean difference was very small ($M_{diff} = .02$). Overall, differences in variability between the grouped and ungrouped rating task were predominantly driven by the trustworthiness condition (see Figure 1).

Figure 1

Rating Variability between Grouped and Ungrouped Rating Tasks across each Trait



Notes. Ungrouped = the rating task where images were presented in randomised order; grouped = the rating task where images were presented in groups consisting of the same identity. A jitter function was applied to reduce overlapping datapoints.

Image Presentation Order and Judgement Variability

Next, we assessed whether trait judgements of faces tended to vary less as participants were exposed to more images of an identity (i.e., after they viewed 10 images of a face). We did not use data from the ungrouped rating task here as we manipulated image order only in Part 2 of the study. We recomputed the variability scores within each image subset, so that the variability scores were calculated separately for images presented in the first and second halves.

We conducted a mixed effects analysis to test whether judgements for images presented earlier in an image set tended to vary more than those presented later. We developed four models. The full model (M4) included trait and image order (first and second 10 images presented in a set) as an interaction term as well as main effects. The reduced model (M3) comprised trait and image order as main effects, whereas the further reduced models M1 & M2 contained single main effects of image order and trait, respectively. All of the models included the same random effects of participant, identity, and image. Implicit beliefs and face gender were not included as random effects due to singularity and convergence issues.

Model Equations:

M4: Variability ~ Trait * Image Order + (1|Participant) + (1|Identity/Image)

M3: Variability ~ Trait + Image Order + (1|Participant) + (1|Identity/Image)

M2: Variability ~ Trait + (1|Participant) + (1|Identity/Image)

M1: Variability ~ Image Order + (1|Participant) + (1|Identity/Image)

We took the same approach of assessing model fit as the previous mixed effects analysis—systematically comparing models with higher-order fixed effects to reduced models. Table 2 displays the fit statistics for each of the models. As seen in Table 2, the Full model demonstrated a significantly better fit to the data than M3, suggesting an interaction between trait condition and the image order. The stimuli random effects (image & identity) in the full model explained only a negligible amount of variance (<.03) in the data, whereas participants accounted for approximately 13% of the variance (*SD* = .37). There was also residual variance in our models (*variance* = .59; *SD* = .77).

Table 2

Model Specifications and Fit Statistics for Image Order and Trait Condition

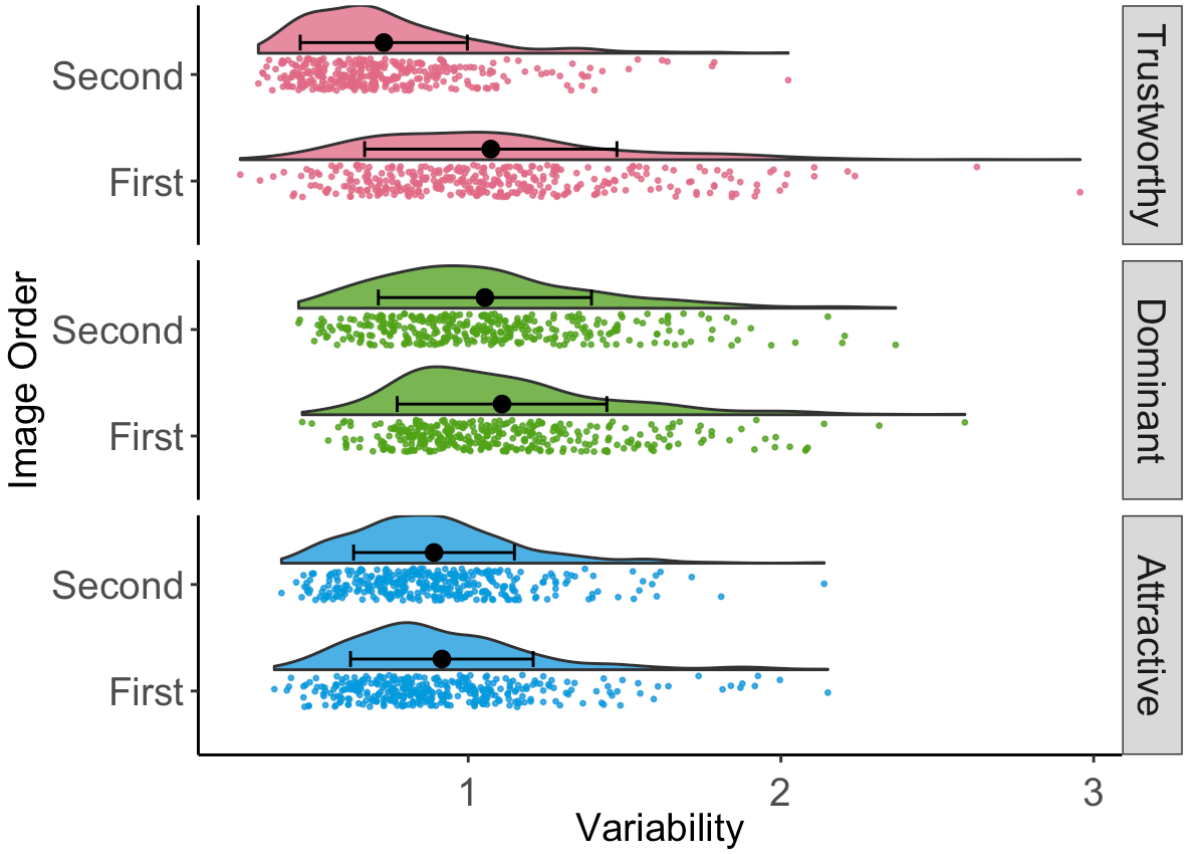
Model Name	Model Specification	Model Fit			Likelihood Ratio Test Statistics			
		AIC	BIC	df	Model Comparison	df	χ^2	*p*-value
M4	Trait X Image Order	79456	79540	10	M3	2	8.15	**.017**
M3	Trait + Image Order	79460	79528	8	-	-	-	-
M2	Trait	79458	79517	7	-	-	-	-
M1	Image Order	79462	79513	6	-	-	-	-

Notes. The likelihood ratio test statistics compare the fit of the higher order models to subsequent reduced models. Participants = 100; observations = 34,000; AIC = Akaike information criterion; BIC = Bayesian information criterion.

We ran pairwise post-hoc tests on model M4 to compare variability in judgements of each trait between the first half and second half of the set. Degrees of freedom were calculated using the Satterthwaite method. As shown in Figure 2, there is some evidence that image presentation order influenced variability of judgements differently across the three traits. Please refer to Table S3 in the Supplemental Materials for descriptive statistics. Ratings of trustworthiness varied significantly more for the first half compared to the second half of the sets $t(29887) = 2.10$, $p = .036$. However, judgements of dominance $t(30229) = .07$, $p = .944$ and attractiveness $t(29800) = 1.93$, $p = .054$ did not differ significantly between images presented earlier and later in image sets. Overall, the interaction between trait and image order was driven only by differences in trustworthiness between the first and second halves of the image sets.

Figure 2

Rating Variability for Images Presented in the First and Second half of Sets across each trait



Notes. First = images presented in the first half of the image set presented to participants; Second = images presented in the second half of the image set presented to participants; a jitter function was also applied to reduce overlapping datapoints.

Beliefs about Personality and Variability in Trait Judgements

Finally, we explored the relationship between implicit person beliefs (Chiu et al., 1997) and the variability of trait judgements from faces. If implicit beliefs about traits are associated with judgement variability in the grouped rating task and not the ungrouped, this might be a mechanism underlying some of the differences in variability between each task. To test Hypothesis 4, we conducted linear mixed effects models separately for each rating task. We opted to run a separate analysis for each rating task as we measured participants' implicit beliefs in both Parts 1 and 2 of the experiment. As such, we wanted to assess whether

participants belief scores obtained in Part 1 were related to variability scores in the ungrouped task and whether belief scores collected in Part 2 were associated with variability in the grouped rating task. The descriptive statistics for the implicit person theory scores in Part 1 (ungrouped rating task; $M = 3.92$; $SD = .83$; $Min = 1.50$; $Max = 6$) and in Part 2 (grouped rating task; $M = 3.86$; $SD = .88$; $Min = 1.62$; $Max = 6$) were similar. Moreover, there was a significant, strong, positive correlation between participant implicit scores in Parts 1 and 2, $r(98) = .88$, $p < .001$, which suggests good test–retest reliability. Although the nature of the experimental manipulation may have influenced the relationship between implicit person scores in each part.

We followed the same analytic approach as the previous mixed effects analyses. We developed the same set of four models to analyse the relationship between implicit beliefs and variability scores for the grouped and ungrouped rating task data. That is, the models listed below are identical for each analysis other than using implicit belief scores that were collected in the corresponding rating task. The full model (M4) contained the interaction of implicit beliefs and trait condition as well as the main effects of each. M3 contained main effects of trait and beliefs, whereas M1 and M2 included single main effects of beliefs and traits, respectively. All four models included the same random effects of participant, image, and identity (face gender and test delay were not included due to singularity issues).

Model Equations:

M4: Variability ~ Beliefs * Trait + (1|Participant) + (1|Identity/Image)

M3: Variability ~ Beliefs + Trait + (1|Participant) + (1|Identity/Image)

M2: Variability ~ Trait + (1|Participant) + (1|Identity/Image)

M1: Variability ~ Beliefs + (1|Participant) + (1|Identity/Image)

Table 3

Model Specifications and Fit Statistics for Implicit Person Scores in the Ungrouped Task

Model Name	Model Specification	Model Fit			Likelihood Ratio Test Statistics			
		AIC	BIC	df	Model Comparison	df	χ^2	*p*-value
M4	Beliefs X Trait	83127	83212	10	M3	2	.52	.771
M3	Beliefs + Trait	83124	83191	8	M2	1	2.30	.129
					M1	2	7.20	**.027**
M2	Trait	83124	83183	7	-	-	-	-
M1	Beliefs	83127	83178	6	-	-	-	-

Notes. The likelihood ratio test statistics compare the fit of the higher order models to subsequent reduced models. Participants = 100; observations = 34,000.

Table 4

Model Specifications and Fit Statistics for Implicit Person Scores in the Grouped Task

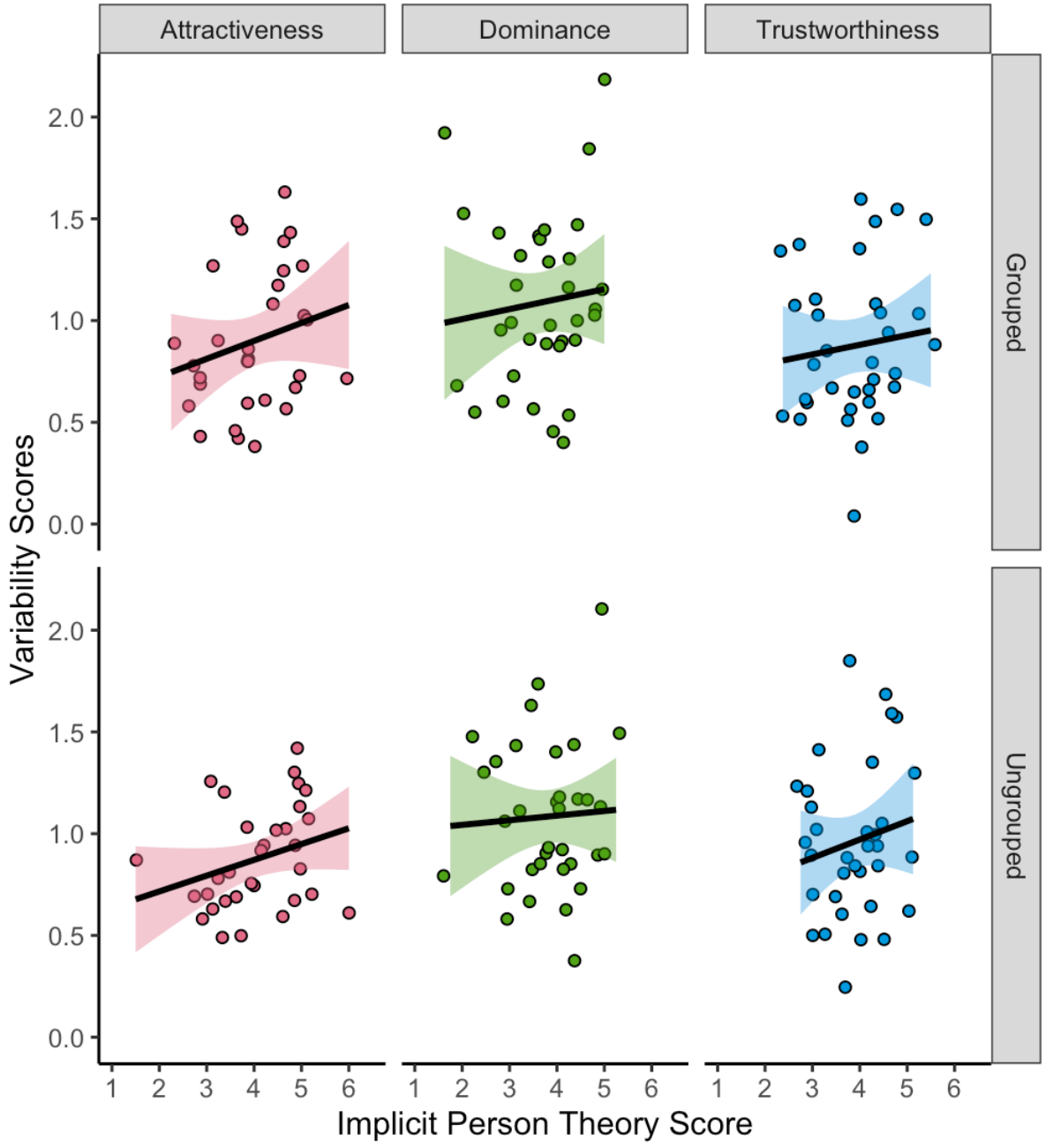
Model Name	Model Specification	Model Fit			Likelihood Ratio Test Statistics			
		AIC	BIC	df	Model Comparison	df	χ^2	*p*-value
M4	Beliefs X Trait	80717	80801	10	M3	2	.18	.913
M3	Beliefs + Trait	80713	80780	8	M2	1	1.93	.164
					M1	2	7.14	**.028**
M2	Trait	80713	80772	7	-	-	-	-
M1	Beliefs	80716	80767	6	-	-	-	-

Notes. The likelihood ratio test statistics compare the fit of the higher order models to subsequent reduced models. Participants = 100; observations = 34,000.

As seen in Tables 3 and 4, the mixed effects models revealed that implicit person beliefs were largely unrelated to trait judgement variability in both the ungrouped and grouped rating tasks. Specifically, in both analyses, the full model containing the interaction of implicit beliefs and trait condition did not show a significantly better fit the data than the reduced model (M3). Similarly, the reduced models which included both main effects did not fit the data significantly better than the single effect of trait (M2) yet did fit the data better than the single effect of beliefs (M1). Overall, these analyses revealed that implicit person belief scores were not meaningfully associated with variability scores in either the ungrouped or grouped rating task. The lack of relationship between implicit beliefs and variability scores for each trait and task is further evidenced in Figure 3. The random effects explained a similar amount of variance in each of the analyses. In the ungrouped rating task analysis, image (*variance* = .01; *SD* = .12), identity (*variance* = .01; *SD* = .10), and participants (*variance* = .11; *SD* = .34) explained very little variance in the data. Similarly, in the grouped rating task, image (*variance* = .01; *SD* = .12), identity (*variance* = .02; *SD* = .15), and participants (*variance* = .15; *SD* = .38). There was residual variance not accounted for by our models in the ungrouped (*variance* = .66; *SD* = .81) and grouped (*variance* = .61; *SD* = .78) rating task analyses.

Figure 3

Implicit Beliefs and Rating Variability for each Task and across each Trait



Notes. The top panel shows the relationships between implicit person theory scores and variability scores in Part 2 (grouped images) and the bottom panel shows Part 1 (ungrouped images). A jitter function was applied to make the overlapping data points more discernible.

Discussion

The aim of this study was to explore whether trait judgements would vary less across images of a face when participants are made aware of the identity. Relatedly, we were interested in assessing whether perceptions of social traits across images of a face would become more consistent after prior exposure to that face. Finally, we explored the role of beliefs about personality in the consistency of trait judgements from faces.

Variability of Trait Judgements

Our first hypothesis that judgements of trustworthiness would vary more across images than dominance, which would vary more than attractiveness was not supported. We found dominance judgements varied significantly more than attractiveness in the ungrouped condition and more than trustworthiness in the grouped condition. Moreover, there were no meaningful differences between the variability of trustworthiness and attractiveness in either task. This was similar to the pattern of results we found in our pilot study. However, our findings conflict with Gogan et al. (2021), who used the same stimuli and same task format as the ungrouped condition. However, judgements of attractiveness consistently varied less across images than trustworthiness and dominance within the current experiment and our pilot study. Trait condition was the only between-subjects manipulation in this experiment; it is possible that our findings might be due to participant differences. There is a growing body of literature suggesting that there is less inter-rater agreement in first impressions from faces than previously thought (Kramer et al., 2018). However, the high ICCs we found in each condition demonstrated good agreement across raters, which suggests participant differences were unlikely to have played a large role in our findings.

Does Knowledge of Identity Change Impressions?

We found some support for our second hypothesis that trait judgements would vary more across images in the ungrouped compared to the grouped rating task. We found

judgements of trustworthiness varied more when images were presented in random order (i.e., ungrouped rating task) than when they were grouped by identity (grouped rating task). However, attractiveness varied more in the grouped than the ungrouped rating task, whereas dominance did not differ significantly between tasks. The interaction between trait and rating condition was primarily driven by the trustworthiness condition.

It is particularly surprising that attractiveness ratings were not more consistent in the grouped compared to the ungrouped condition given that Goller et al. (2018) and Richie et al. (2017) found perceptions of attractiveness to be anchored to past judgements. Presumably, it would be easier for participants to use previous judgements of a face as a reference point for a current judgement when the identity of the images is known (i.e., the grouped condition of our study). Our findings suggest that being aware of which images belong to a given identity might attenuate perceptions of trustworthiness across different instances of a face. Perhaps it is more important to use past judgments of trustworthiness to inform future decisions than other traits. For instance, if someone is untrustworthy in one setting, it might be wise to be wary of them in future encounters.

Perceptual Familiarity and Trait Judgements

Building upon the previous hypothesis, it is of theoretical interest to assess whether trait judgements stabilise as one becomes perceptually familiar with an identity. A hallmark of familiar face identification is the ability to discount image variability (Andrews et al., 2015); however, it is unclear whether this effect extends to traits. We found some support for our third hypothesis that trait judgements of the first half of images in a given identity presented in the grouped rating task would vary more than the subsequent second half of images. However, only judgements of trustworthiness varied significantly more for images presented in the first compared to the second half of sets, whereas dominance and attractiveness did not differ significantly.

It is possible that perceptual familiarity alone might have a small effect on stabilising trait judgements but have a larger effect when semantic knowledge is also available. Moreover, our findings suggest that perceptual familiarity does not stabilise perceptions of social traits to the same extent as perceptions of identity (Jenkins et al., 2011). Jenkins et al. (2011) found that almost all participants were able to complete a face sorting task (a test of face identification abilities) with 100% accuracy when they were familiar with the identities. This suggests that familiarity almost completely negates the interference of image variability when processing the identity of a face. However, identity tasks entail binary judgments (same or different identity), whereas trait judgments are more continuous. A better comparison may be to examine how familiarity affects ratings of image likeness. There is evidence that familiarity increases the perceived likeness of images (Ritchie et al., 2018), suggesting that familiarity reduces perceived variability in identity.

A key difference between the current study and past research (e.g., Jenkins et al., 2011; Mileva et al., 2019) is the use of celebrities. Interestingly, Wiese et al. (2021) recently found that personally familiar faces produced similar event-related potentials to faces of celebrities, suggesting that different types of familiarity do not differ qualitatively. However, it is difficult to control for the degree (i.e., quantity) of past exposure to celebrity faces. Potentially, we might have found stronger effects of perceptual familiarity in our study if participants were exposed to more images of each identity. However, it is worth noting that the familiarity induced by the presentation order of the images in the grouped rating task should have been enhanced further due to prior exposure to the images in the ungrouped rating task. Nonetheless, future researchers are encouraged to further explore the difference between perceptual and semantic familiarity when forming impressions, while controlling for past exposure.

Beliefs about Traits and Variability in Trait Judgements

We did not find support for our final hypothesis that the tendency to believe personality attributes are malleable (rather than fixed) would be associated with greater variability in trait judgements of images. Our analyses revealed that participants who scored high on the implicit person theory scale (i.e., believed personality is not fixed) did not provide more variable trait judgements across images in either the ungrouped or the grouped rating tasks. Similarly, including beliefs as a random effect did not explain much variance when testing the effect of knowledge of identity on perceptions of traits. Our pilot study revealed comparatively stronger relationships between trait ratings and participant beliefs. However, in the pilot study we measured participant beliefs after completing the rating task. Taken together, our findings might suggest that exposure to variable images amplify beliefs about personality, as Weisbuch et al. (2016) proposed, more than these beliefs guide trait impressions. Therefore, these higher-order beliefs do not readily explain any effects of knowledge of identity on trait variability.

Limitations

Some limitations of the current study should be noted. First, as with all face perception research, our conclusions are inherently limited by the specific images used in this study. Given the variations in appearance both between and within identities, it is possible that our results would change with the use of different stimuli (e.g., different images of the same identities and/or photos of different identities). We mitigated this limitation by drawing upon a large database containing numerous images of several identities.

Second, the order of the grouped and the ungrouped rating task was not randomised. All participants completed the ungrouped rating task first in Part 1 and then the grouped rating task in Part 2. It is possible that this might have led to practice effects or induced some familiarity with the images. In the pilot study we employed a between-subjects design;

however, it is likely that participant differences overshadowed any effect of the different rating tasks. We opted to not randomise the order of the rating tasks in the current study because completing the grouped rating task first could have confounded results in the ungrouped rating task. If participants completed the grouped rating task first, it is likely that they would have learned which images belong to a given identity when completing the ungrouped rating task.

Third, a potential explanation for why we did not find more pronounced differences between images presented in the first compared to the second half of within an image set is that 10 images might not be enough to induce familiarity. Becoming familiar with a face is a graded process (Clutterbuck & Johnston, 2002). It is hard to say at what point participants would have become perceptually familiar with a given identity, if at all. Future research may wish to use more sensitive measures of familiarity to assess how it influences variability of trait judgements.

Finally, we were not able to entirely control for the influence of top-down contextual information in our study. Specifically, in the grouped rating task, participants were informed that each set would contain images of a single identity which were given a fake name. Andrews et al. (2015) demonstrated that informing participants of the number of identities present in a face sorting task enhanced performance. Similarly, Honig et al. (2021) showed that merely labelling images with a name influenced face identification. Given the potential influence of top-down processes, it is unclear how much of our findings are due to perceptual familiarity alone. Relatedly, the current study design makes it difficult to disentangle what is driving the reduced variability of trustworthiness judgements in the grouped task. Namely, it is not clear whether judgement variability was reduced due to making participants aware of the identity or induced from prior familiarity with the images in the ungrouped rating task. Although we attempted to mitigate the influence of prior exposure to the images by including

a test delay between the two tasks, some perceptual familiarity effects might have played a role in our findings. Future researchers are encouraged to employ a between-subject design to compare variability in trait judgements for the grouped and ungrouped tasks, thereby isolating the effect of knowledge of identity without being confounded by prior familiarity.

Conclusion

We embarked on this research inspired by recent claims that trait impressions vary as much or more within identities as they do between identities, suggesting that trait judgments have little to do with face identity and everything to do with image properties. In our view, such claims are hard to reconcile with experience, and with our own research which has shown some traits vary more than others and that differences between identities persist despite substantial within-person variability. In the study reported here we sought to examine whether image-related within-person variation in trait judgments is constrained by identity, and if so, whether this is attributable to a top-down influence of beliefs about how traits vary within people, or due to bottom-up perceptual processes. Overall, our results suggest that knowledge of identity may constrain impressions to some degree, and there was some evidence that impressions stabilise with perceptual familiarity. However, knowledge of identity and perceptual familiarity appeared to take effect only for trustworthiness judgements but not dominance or attractiveness. We found that impressions about how trustworthy, dominant, or attractive people appeared in different images were largely unrelated to implicit person beliefs and did not readily explain why knowledge of identity constrained variability. Thus, any effect of knowledge of identity or familiarity on trait variability is probably more about perceptual processes than higher-order beliefs. This, together with the research of Gogan et al. (2021), suggests to us that trait impressions are not entirely about image properties but also linked to relatively stable variations that become

incorporated into abstract face representations. Unfortunately for most of us, a good picture will probably not make us as attractive as Brad Bitt or Jennifer Aniston.

Acknowledgements

This research was supported by the Australian Government Research Training Program Scholarship to the first author.

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/rqjcx/

Declaration of conflict of interest

We have no conflicts of interest to declare.

References

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology*, *68*(10), 2041-2050. https://doi.org/10.1080%2F17470218.2014.1003949

Armann, R. G., Jenkins, R., & Burton, A. M. (2016). A familiarity disadvantage for remembering specific images of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *42*, 571–580. http://dx.doi.org/10.1037/xhp0000174

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, *73*(1), 105-116. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

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

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1467-1485. doi:10.1068/p3335. https://doi.org/10.1080/17470218.2013.800125

Burton, A. M., Jenkins, R., & Schweinberger, S. R. (2011). Mental representations of familiar faces. *British Journal of Psychology*, *102*(4), 943-958. https://doi.org/10.1111/j.2044-8295.2011.02039.x

Calder, A. J., Young, A. W., Keane, J., & Dean, M. (2000). Configural information in facial expression perception. *Journal of Experimental Psychology: Human perception and performance*, *26*(2), 527. https://psycnet.apa.org/doi/10.1037/0096-1523.26.2.527

Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception*, *31*(8), 985-994. https://doi.org/10.1068%2Fp3335

Chiu, C. Y., Hong, Y. Y., & Dweck, C. S. (1997). Lay dispositionism and implicit theories of personality. *Journal of Personality and Social Psychology*, *73*(1), 19. https://psycnet.apa.org/doi/10.1037/0022-3514.73.1.19

Dweck, C. S., Chiu, C, & Hong, Y. (1995). Implicit theories and their role in judgments and reactions: A world from two perspectives. *Psychological Inquiry, 6.* 267-285. https://doi.org/10.1207/s15327965pli0604_1

Gogan, T., Beaudry, J., & Oldmeadow, J. (2021). Within-Person Variability in First Impressions from Faces. *Perception*, *50*(7), 595–614. https://doi-org.ezproxy.lib.swin.edu.au/10.1177/03010066211019727

Goller, J., Leder, H., Cursiter, H., & Jenkinlavas, R. (2018). Anchoring Effects in Facial Attractiveness. *Perception*, *47*(10-11), 1043-1053. https://doi.org/10.1177%2F0301006618802696

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

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

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

Kramer, R. S. (2016). Within-person variability in men's facial width-to-height ratio. PeerJ, 4,. https://doi.org/10.7717/peerj.1801

Kramer, R. S. S., Mileva, M., & Ritchie, K. L. (2018). Inter-rater agreement in trait judgements from faces. *PLoS One, 13,* 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. https://doi.org/10.1037/0033-2909.110.

Lavan, N., Mileva, M., Burton, A. M., Young, A. W., & McGettigan, C. (2021). Trait evaluations of faces and voices: Comparing within-and between-person variability. *Journal of Experimental Psychology: General*.

Lenth, R. (2019). emmeans: Estimated marginal means, aka least-squares means (R package, Version 1.4) [Computer software]. https://CRAN.Rproject.org/package=emmeans

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

Penton-Voak, I., & Perrett, D. I. (2000). Consistency and individual differences in facial attractiveness judgements: An evolutionary perspective. *Social Research*, 219-244.

Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version =1.8.12.

Ritchie, K. L., Kramer, R. S., & Burton, A. M. (2018). What makes a face photo a 'good likeness'?. *Cognition*, *170*, 1-8. https://doi.org/10.1016/j.cognition.2017.09.001

Ritchie, K. L., Palermo, R., & Rhodes, G. (2017). Forming impressions of facial attractiveness is mandatory. *Scientific Reports*, *7*(1), 1-8. https://doi.org/10.1038/s41598-017-00526-9

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology*, *57*, 199–226. https://doi.org/10.1146/annurev.psych.57.102904.190208

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.

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., & Rhodes, G. (2017). Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties. *British Journal of Psychology*, *108*(2), 397-415. https://doi.org/10.1111/bjop.12206

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*, 1623- 1626. https://doi.org/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. https://doi.org/10.1146/annurev-psych-113011-143831

Vernon, R. J., Sutherland, C. A., Young, A. W., & Hartley, T. (2014). Modelling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, *111*(32), E3353-E3361. https://doi.org/10.1073/pnas.1409860111

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192-196. https://doi.org/10.3758/BF03206482

Walsh, V., & Kulikowski, J. (Eds.). (1998). *Perceptual constancy: Why things look as they do*. Cambridge University Press.

Weisbuch, M., Grunberg, R. L., Slepian, M. L., & Ambady, N. (2016). Perceptions of variability in facial emotion influence beliefs about the stability of psychological characteristics. *Emotion*, *16*(7), 957-964. https://psycnet.apa.org/doi/10.1037/emo0000123

Westfall, J., Nichols, T. E., & Yarkoni, T. (2016). Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *Welcome Open Research*, *1*. https://doi.org/10.12688/wellcomeopenres.10298.2

Wiese, H., Hobden, G., Siilbek, E., Martignac, V., Flack, T. R., Ritchie, K. L., ... & Burton, A. M. (2021). Familiarity is familiarity is familiarity: Event-related brain potentials reveal qualitatively similar representations of personally familiar and famous faces. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://psycnet.apa.org/doi/10.1037/xlm0001063

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686, https://doi.org/10.21105/joss.01686

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

Winter, B. (2013). A very basic tutorial for performing linear mixed effects analyses. *arXiv preprint arXiv:1308.5499.* http://arxiv.org/pdf/1308.5499

White, D., Sutherland, C. A., & Burton, A. L. (2017). Choosing face: The curse of self in profile image selection. *Cognitive Research: Principles and Implications*, *2*(1), 1-9. https://doi.org/10.1186/s41235-017-0058-3

Young, A. W., McWeeny, K. H., Hay, D. C., & Ellis, A. W. (1986). Matching familiar and unfamiliar faces on identity and expression. *Psychological Research*, *48*(2), 63-68. https://doi.org/10.1007/BF00309318

Appendix A

Deviations from Preregistration

We made some deviations to our preregistration. First, our preregistration outlined that we expected to recruit approximately 117 participants, yet we collected data from 126 due to underestimating the number of data exclusions. Although not preregistered, we calculated ICCs for each condition in order to assess rater agreement and whether participant differences accounted for differences in findings between conditions. We also revised our model comparison approach for the mixed effects analyses, which originally was to compare each model to the null model. Our revised approach consisted of comparing higher-order models to reduced models in order to account for lower-order effects in each comparison. Although not preregistered, we performed post-hoc tests for our mixed effects models in order to explore significant interactions more thoroughly. We removed target gender as a fixed effect in the first mixed effects analysis as gender was not relevant to our hypotheses nor a key focus of this paper. However, we included gender as a random effect given its importance in first impressions (Mileva et al., 2019). Finally, we employed mixed effects models to assess the influence of implicit person beliefs instead of Pearson correlations to account for participant differences (added as a random effect) and the hierarchical structure of the data. For completeness, we report our original preregistered analyses in the supplemental materials (see Supplement 4).

Appendix B

Implicit Person Theory Scale Items (Chiu et al., 1997)

Items with "*" were reverse scored

1. The kind of person someone is, is something very basic about them and it can't be changed very much.
2. People can do things differently, but the important parts of who they are can't really be changed.
3. *Everyone, no matter who they are, can significantly change their basic characteristics.
4. As much as I hate to admit it, you can't teach an old dog new tricks. People can't really change their deepest attributes.
5. *People can always substantially change the kind of person they are.
6. Everyone is a certain kind of person, and there is not much that can be done to really change that.
7. *No matter what kind of person someone is, they can always change very much.
8. *All people can change even their most basic qualities.