

Static and Dynamic Facial Cues Differentially Affect the Consistency of Social Evaluations

Personality and Social
Psychology Bulletin
1–12
© 2015 by the Society for Personality
and Social Psychology, Inc
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146167215591495
pspb.sagepub.com


Eric Hehman¹, Jessica K. Flake², and Jonathan B. Freeman¹

Abstract

Individuals are quite sensitive to others' appearance cues when forming social evaluations. Cues such as facial emotional resemblance are based on facial musculature and thus dynamic. Cues such as a face's structure are based on the underlying bone and are thus relatively static. The current research examines the distinction between these types of facial cues by investigating the consistency in social evaluations arising from dynamic versus static cues. Specifically, across four studies using real faces, digitally generated faces, and downstream behavioral decisions, we demonstrate that social evaluations based on dynamic cues, such as intentions, have greater variability across multiple presentations of the same identity than do social evaluations based on static cues, such as ability. Thus, although evaluations of intentions vary considerably across different instances of a target's face, evaluations of ability are relatively fixed. The findings highlight the role of facial cues' consistency in the stability of social evaluations.

Keywords

social evaluation, face perception, impression formation, non-verbal cues

Received January 26, 2015; revision accepted May 22, 2015

The cues contained in others' faces evoke powerful effects on how we perceive and interact with them (Brewer, 1988; Fiske & Neuberg, 1990). Perceivers often largely agree in the inferences made about others from their appearance (Berry, 1991; Kenny & Albright, 1987; Moskowitz, 1990), and these inferences can be consequential, predicting real-world outcomes including political success, financial performance, and judicial decisions (Hehman, Carpinella, Johnson, Leitner, & Freeman, 2014; Todorov, Mandisodza, Goren, & Hall, 2005; Wong, Ormiston, & Haselhuhn, 2011; Zebrowitz & McDonald, 1991).

Given the importance of these inferences, understanding the underlying facial cues from which they arise is critical. Researchers as far back as Darwin (1872) have postulated that human facial musculature evolved to communicate emotional information to other humans. Accordingly, we are quite sensitive to even subtle emotional resemblances in a face, such that they are overgeneralized to infer another's personality characteristics (Zebrowitz, Andreoletti, Collins, Lee, & Blumenthal, 1998). That is, although a face may appear emotionally "neutral," slight resemblances to emotional expressions through natural variations in facial morphology or temporary muscle contractions facilitate corresponding trait inferences. For instance, a man with temporarily downward turned brows, perceived in that instant, is regarded as more ill-tempered due to his face's resemblance

with anger expressions. Because targets with directed expressions of anger likely have negative intentions toward the perceiver, the faces of individuals who slightly resemble these expressions are overgeneralized to be perceived as less likeable or trustworthy, whereas those with slightly happy expressions are perceived as having more positive intentions (Oosterhof & Todorov, 2008; Zebrowitz, Fellous, Mignault, & Andreoletti, 2003).

Many studies have found that perceivers reach strong consensus in their assessments of intentions (e.g., trustworthy–untrustworthy, good–bad) from a single facial photo, and these assessments are driven largely by the emotion overgeneralization effects described above. This dimension in social evaluation has been variously called Warmth, Valence, Basic Trust, Need for Tenderness, and Trustworthiness, among many others. To avoid confusion, henceforth we refer to this dimension as intentions. Intriguingly, recent work has found that across multiple photos of a single individual, evaluations

¹New York University, New York City, USA

²University of Connecticut, Storrs, USA

Corresponding Author:

Eric Hehman, Department of Psychology, New York University, 6 Washington Place, New York University, New York, NY 10003, USA.
Email: eric1hehman@gmail.com

of intentions can be considerably unstable. Because different photos of an individual can contain different emotional resemblances, perceived intentions from each photo could vary considerably due to overgeneralization effects (Rule, Krendl, Ivcevic, & Ambady, 2013; Todorov & Porter, 2014). In Kenny's (1991) Consensus–Accuracy model, one parameter critical to consensus is consistency, or the stability of a target's appearance or behavior across multiple presentations. Consistent with this model, because evaluations of intentions may vary from moment to moment as they tend to be based on ephemeral emotional resemblances from dynamic facial musculature (Todorov & Porter, 2014), social evaluations arising from these cues might demonstrate relatively low stability.

Although intentions may be a powerful dimension underlying our social evaluations (Erikson, 1950; Freedman, Leary, Ossario, & Coffey, 1953; Oosterhof & Todorov, 2008), there are other important dimensions that arise from unique facial cues, such as physical ability. Rapidly assessing ability and potential threat would be quite adaptive, given that the lethal intergroup competition persisting throughout human history is theorized to have had a significant impact on our evolutionary development (McDonald, Asher, Kerr, & Navarrete, 2011; Tooby & Cosmides, 1988; Van Vugt, 2009). Critically, cues related to evaluations of ability tend to be relatively static and structural rather than dynamic and malleable. Thus, in accordance with the Consensus–Accuracy model (Kenny, 1991), these cues would demonstrate a higher level of consistency, varying less across multiple photos of an individual. For instance, the facial width-to-height ratio (fWHR) is a facial metric central to evaluations of ability (Carré, McCormick, & Mondloch, 2009; Carré, Morrissey, Mondloch, & McCormick, 2010; Hehman, Leitner, Deegan, & Gaertner, 2015; Hehman, Leitner, & Gaertner, 2013). It is a static cue derived from the underlying bone structure: a face's bizygomatic width (i.e., the distance between the left and right zygion) divided by the upper cranial facial height (i.e., the distance between the upper lip and mid-brow; Figure 1). Testosterone has been proposed as a link between fWHR and behavior (Carré & McCormick, 2008), and recent work has demonstrated that high fWHR males have higher levels of circulating and reactive testosterone (Lefevre, Lewis, Perrett, & Penke, 2013). Indeed, a high testosterone-to-estrogen ratio is thought to specifically facilitate the lateral growth of the cheekbones, mandibles, chin, and the forward growth of the bones of the eyebrow ridges (Farkas, 1994).

Importantly, research indicates perceivers are highly sensitive to this cue when making evaluations related to power and ability. High fWHR males are perceived as more intimidating, stronger, and more aggressive in a variety of contexts (Carré et al., 2009; Hehman, Leitner, Deegan, & Gaertner, 2013; Hehman, Leitner, & Freeman, 2014). Furthermore, research examining the basis for these evaluations found that fWHR, and not other related facial cues, was uniquely responsible for these evaluations (Carré et al., 2010).

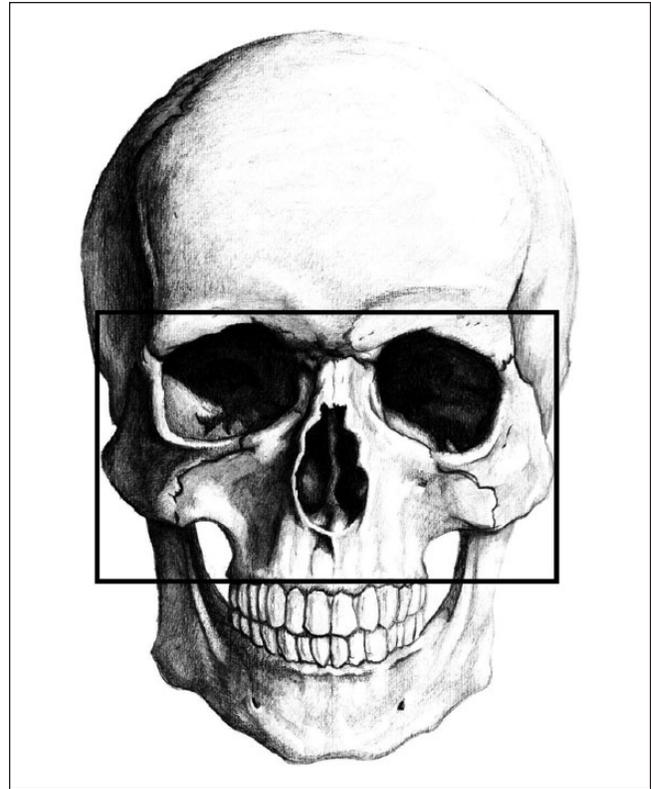


Figure 1. Example facial width-to-height coding based on the underlying bone structure.

Thus, evaluations of intentions tend to be based on dynamic, malleable facial musculature that can vary across multiple instances of an individual. In contrast, evaluations such as ability tend to be based on more static facial cues that arise from underlying bone structure, less likely to change across multiple instances. Because consistency of a trait is an important predictor of consensus (Ambady, Bernieri, & Richeson, 2000; Kenny, 1991), we hypothesize that evaluations of intentions will be considerably less consistent across multiple instances of a target relative to evaluations of ability. Furthermore, we hypothesize that this discrepancy will be due to the malleability of dynamic cues (e.g., emotional resemblances) driving perceived intentions versus the stability of static cues (e.g., fWHR) driving perceived ability. In other words, here we examine whether individuals might be able to readily change how well-intentioned they are perceived in a photo, but be relatively unable to change the ability conveyed. Testing this possibility is also important for the growing number of researchers examining face-based social evaluations, assessing to what extent a single photo might be used as a reasonable representation of an individual. We test these hypotheses in the following four studies.

Study 1

Here we tested whether social evaluations derived from static cues (i.e., ability) are more consistent across multiple

portrayals of single individuals, whereas evaluations derived from dynamic cues (i.e., intentions) are malleable and less consistent. Although “dynamic” frequently refers to video-based stimuli in the impression formation literature, throughout the current work, we use photographic stimuli. Here we use the terms “static” and “dynamic” to describe the extent to which specific cues can be static or dynamic in real-world settings.

Method

Participants. 119 participants (69 female, Age $M = 33$) rated faces through Amazon’s Mechanical Turk for monetary compensation.

Stimuli. The same male stimuli used in recent work finding malleability in perceived intentions (Todorov & Porter, 2014) were downloaded from the FERET database. This database was constructed to test face- and identity-detection algorithms, and so the multiple photos of the same individuals purposely varied in features (e.g., emotional expression), although this was not expressly manipulated (for greater detail, see Phillips, Wechsler, Huang, & Rauss, 1998). In total, we presented 5 photos of 10 different identities (50 faces) that varied in ethnicity (7 White, 2 Asian, 1 Middle Eastern).

Procedure. Participants rated each face, presented in random order, on a 1 (*not at all [trait]*) to 7 (*very [trait]*) scale for only one randomly assigned trait: Friendly, Trustworthy, or Physically Strong. At least 38 participants made ratings on each trait. After the ratings, demographic information was collected.

Normed ratings and coding. We tested whether emotional resemblance had greater variability than facial structure. Prior research has found for emotionally neutral faces that subtle resemblance to an angry expression is associated with perceived negative intentions (e.g., untrustworthy), and resemblance to a happy expression is associated with perceived positive intentions (e.g., trustworthy; Engell, Haxby, & Todorov, 2007; Oosterhof & Todorov, 2009). To assess such emotional resemblance, a separate group of participants on Mechanical Turk ($n = 41$) rated the faces on emotion, rather than trait attributions, using a 1 “very angry” to 7 “very happy” scale. Ratings were averaged across participants. To assess facial structure, three raters blind to the hypotheses coded each face for fWHR based on anatomical points (see Figure 1) defined by previous research (Carré & McCormick, 2008; Carré et al., 2010). Stimuli ($n = 3$) that were too rotated to measure accurately were not included. Inter-rater reliability was high ($\alpha = .90$), and ratings were averaged.

Data preprocessing and analytic approach. Participants who either made ratings in less than 400 ms or made the same rating

across the majority of trials were removed from analysis ($n = 3$, 3% of the data). Remaining ratings were then averaged for each individual face. As expected, friendliness and trustworthiness were highly correlated ($r = .715$, $p < .001$, 99.5% confidence interval (CI) = [0.474, 0.882]) and averaged to form an intentions dimension. The face identity acted as the unit of analysis in multilevel models in which different faces of the same identity were nested within identity, thereby testing our hypotheses while partialling out between-identity variance.

Results

We compared the variance in ratings between intentions and ability by modeling heterogeneity of Level 1 variance in a multilevel framework, as greater variance would indicate more variability in ratings across multiple photos of a single individual. In this analysis, a log linear regression is used to model the Level 1 variance, σ^2 , such that $\text{Log}(\sigma^2) = \alpha_0 + \alpha_1$, where the Level 1 variance is transformed into a natural log and modeled as a function of rating type. Intentions ratings were coded 0, and ability ratings were coded 1. Thus, α_0 is the log of the variance for ratings of intentions, and α_1 is the difference in the log of the variance for ratings of ability. The difference in the variance between the two types of ratings was significant, $\alpha_1 = -3.15$, $z = -10.42$, $p < .001$. The predicted variance for intentions ratings was .90, whereas the predicted variance for the ability ratings was .04. Accordingly, the heterogeneous model, which allows intentions and ability to have different variances, better fit the data than the homogeneous model, $\Delta\chi^2 = 32.10$, $\Delta df = 1$, $p < .001$.¹ Consistent with our predictions, these results indicate that evaluations of ability had less variance than evaluations of intentions (Figure 2).

Given this result, we would similarly expect emotional resemblances to exhibit greater variance than fWHR (as it is the variant vs. invariant nature of these cues that may produce more consistent evaluations of intention vs. less consistent evaluations of ability). Results from another model of heterogeneity of this Level 1 variance were consistent with this interpretation: The difference in the variance between the two types of ratings was significant, $\alpha_1 = -1.39$, $z = -4.439$, $p < .001$, and the heterogeneous model again fit the data significantly better than the homogeneous model, $\Delta\chi^2 = 6.76$, $\Delta df = 1$, $p = .009$. These results support our claim that social evaluations derived from static facial structure are more consistent than those derived from dynamic cues. Study 2 builds on these results by testing our hypotheses more rigorously using computer-generated faces.

Study 2

Whereas we measured emotional resemblance and facial structure in Study 1, Study 2 manipulated them. Computer-generated faces independently varied facial cues across a single target identity, affording greater precision and control.

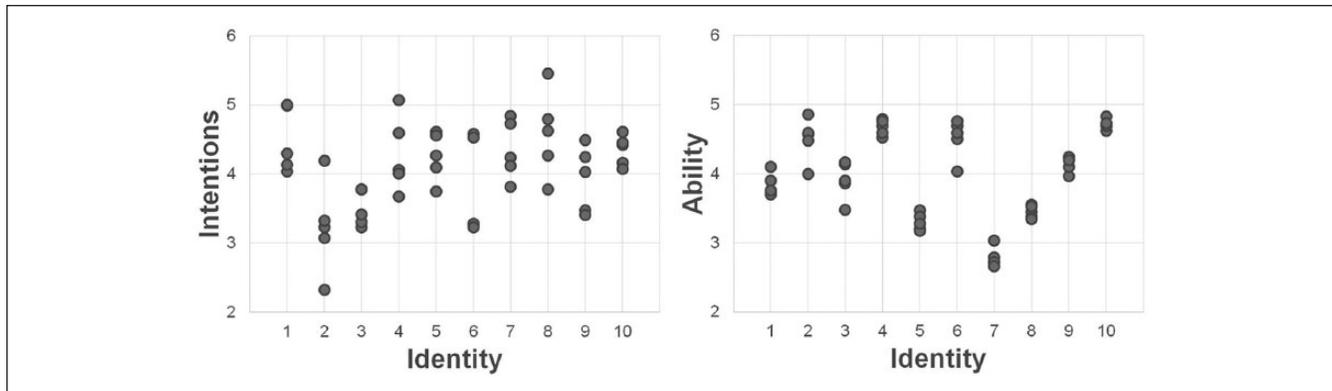


Figure 2. The ratings for each stimulus face across all ten identities in Study 1, demonstrating the greater variability in ratings of intentions than of ability.

As in Study 1, we predicted that social evaluations derived from dynamic cues (i.e., intentions) would be less consistent than social evaluations derived from static cues (i.e., ability).

Method

Participants. Four hundred and forty-seven participants (214 female, Age $M = 38$) rated faces through Mechanical Turk for monetary compensation.

Stimuli. Forty total identities were generated using FaceGen Modeler (www.facegen.com) to independently manipulate emotional resemblance, and facial structure was manipulated by increasing or decreasing the aspect ratio of each image as in previous research (Hehman et al., 2015; Hehman, Leitner, & Freeman, 2014). FaceGen creates 3D digital models derived from a database of laser scans of human faces (Banz & Vetter, 1999). These faces can be morphed to appear with various emotional expressions. For each identity, emotional resemblance was manipulated on a 5-point continuum ranging from slightly resembling an angry expression to slightly resembling a happy expression, although all faces were ostensibly emotionally neutral² (Figure 3C). fWHR was manipulated on a 4-point continuum, resulting in 20 faces within each identity, or 800 total faces.

Procedure. Participants rated each face on items and in a procedure identical to Study 1, except that ratings of Warmth were additionally collected to increase reliability.

Data preprocessing and analytic approach. As in Study 1, 71 participants (16% of the data) who either made ratings in less than 400 ms or made the same rating across the majority of trials (>80%) were removed from analysis. Remaining ratings were then averaged for each individual face. Trustworthiness, Warmth, and Friendliness were highly related ($\alpha = .951$) and

averaged to form a composite score of intentions. As in Study 1, the identity again acted as the level of analysis in a series of multilevel models, thereby testing our hypotheses while partialling out between-identity variance.

Results

Manipulation checks. To confirm that the manipulation of emotion resemblance (angry–happy) influenced perceived intentions and manipulation of facial structure (fWHR) influenced perceived ability, we regressed the ratings onto these two dimensions. As expected, emotional resemblance was predictive of intentions ($\gamma_{20} = 2.228$, $SE = .079$, $t = 28.30$, $p < .001$): Targets with a stronger resemblance to a happy expression were perceived with more positive intentions (Figure 3A). Similarly, facial structure was predictive of ability ($\gamma_{10} = 1.100$, $SE = .137$, $t = 8.01$, $p < .001$): Targets with wider facial structure were rated as higher in ability (Figure 3B). Interestingly, however, these dimensions were not completely independent. Faces with wider facial structure were additionally rated as having more positive intentions ($\gamma_{20} = .483$, $SE = .094$, $t = 5.12$, $p < .001$), and faces with slightly angrier resemblances were rated as higher in ability ($\gamma_{10} = -.019$, $SE = .077$, $t = -2.55$, $p = .015$). We discuss the non-independence of these dimensions later.

If facial structure is primarily driving evaluations of ability, visual obscuration of the apparent width of the face should disrupt the evaluation of this trait. Further tests confirmed this possibility by cropping the external width of each face. Additional participants ($n = 131$, 70 female, Age $M = 43$) rated each face on identical items. Only variants of 10 identities were presented for expediency (200 total faces). Supportive of our hypotheses, although ratings of intentions continued to track emotional resemblances ($\gamma_{20} = 2.231$, $SE = .196$, $t = 11.41$, $p < .001$), ratings of ability were entirely unrelated to the obscured facial width manipulation ($\gamma_{10} = -.246$, $SE = .530$, $t = -.46$, $p = .643$).

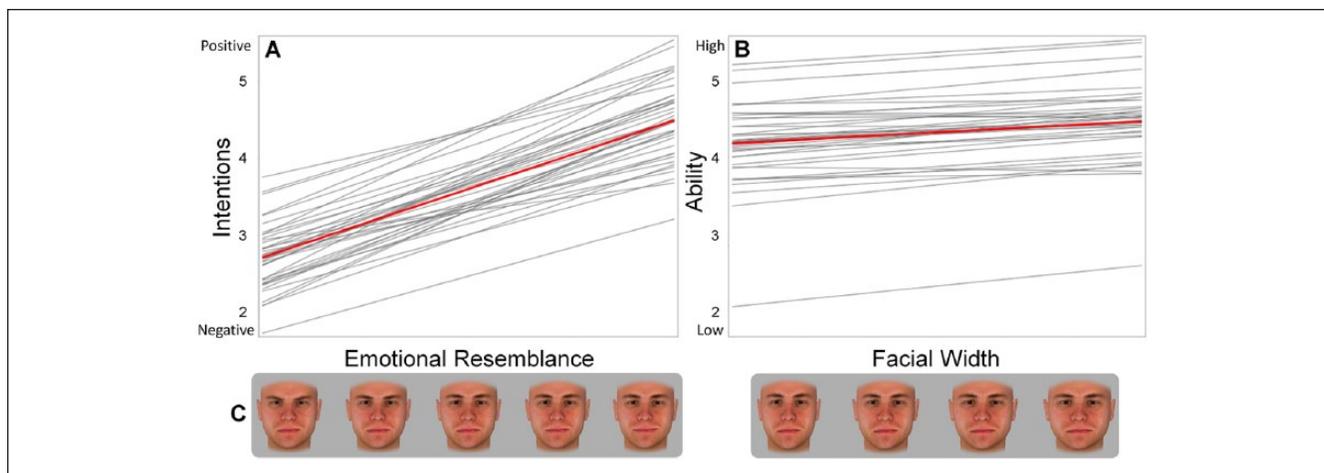


Figure 3. Correlations between stimulus manipulations and ratings hypothesized to be most sensitive in Study 2: (A) The impact of the emotional resemblance manipulation on intentions ratings; (B) The impact of the facial width manipulation on ability ratings; (C) Example of the stimuli manipulations in Study 2.

Note. Thin gray lines indicate results for each target identity, whereas the thicker red (dark gray) line represents the grand slope across all target identities.

Identically, should emotional resemblance be responsible for evaluating another's intentions, visual obscuration of the area around the eyes and mouth where relevant muscles (e.g., zygomaticus, orbicularis occuli) are located should disrupt the relationship between emotional resemblance and ratings of intentions. Accordingly, the internal facial features of the stimuli were obscured. Additional participants ($n = 121$, 47 female, Age $M = 31$) rated each face on identical items. Again, variants of 10 identities were presented for expediency (200 total faces). Although ratings of ability continued to track facial width ($\gamma_{10} = 2.490$, $SE = 1.090$, $t = 2.28$, $p = .048$), the obscured emotional resemblance manipulation was no longer predictive of intentions ($\gamma_{20} = -.225$, $SE = .494$, $t = -.46$, $p = .649$). Together, these results strongly support our manipulations, indicating that emotional resemblance primarily drives perceived intentions and facial structure drives perceived ability.

Variance analysis. To address our primary hypothesis, we compared the variance in ratings of intentions and ability by modeling heterogeneity of Level 1 variance in a multilevel framework, as in Study 1. Supporting our hypothesis, the difference in the log of the variance between the two types of ratings was significant, $\alpha_1 = -1.23$, $z = -16.77$, $p < .001$. The predicted variance for intentions ratings was .79, whereas the predicted variance for the ability ratings was .23. The heterogeneous model, which allows intentions and ability to have different variances, again better fit the data than the homogeneous model, $\Delta\chi^2 = 110.71$, $\Delta df = 1$, $p < .001$ (Figure 4). These results extend the findings of Study 1 to precisely manipulated facial images, providing further support for the claim that evaluations of ability are less variable across multiple instances of a target than evaluations of intentions.

Study 3

Here, we examine to what extent the relative malleability in perceived intentions versus stability in perceived ability (found in Studies 1 and 2) would manifest in downstream social decisions. Participants selected suitable targets in a context in which either perceived intentions or ability should have driven selection decisions. We hypothesized that selection decisions primarily related to intentions would exhibit more variability across multiple instances of a target relative to selection decisions primarily related to ability.

Method

Participants. One hundred and three (56 female) participants selected faces from an array through Mechanical Turk for monetary compensation.

Stimuli. Participants were randomly presented with one of four arrays. In each array, eight identities were displayed in two rows (Figure 5). Across the four arrays, four variations of each identity were present: a subtly angry-resembling high- and low-fWHR version of the identity, and a subtly happy-resembling high- and low-fWHR version (which identity appeared in what condition varied across each slide). Modeled off previous research (Hehman et al., 2015), this approach introduces variability in emotional resemblance and facial structure while controlling for individual differences in appearance between each target.

Procedure. Participants were tasked with selecting four of the faces under one of two conditions. In the Financial Advisor condition, participants read a short paragraph informing

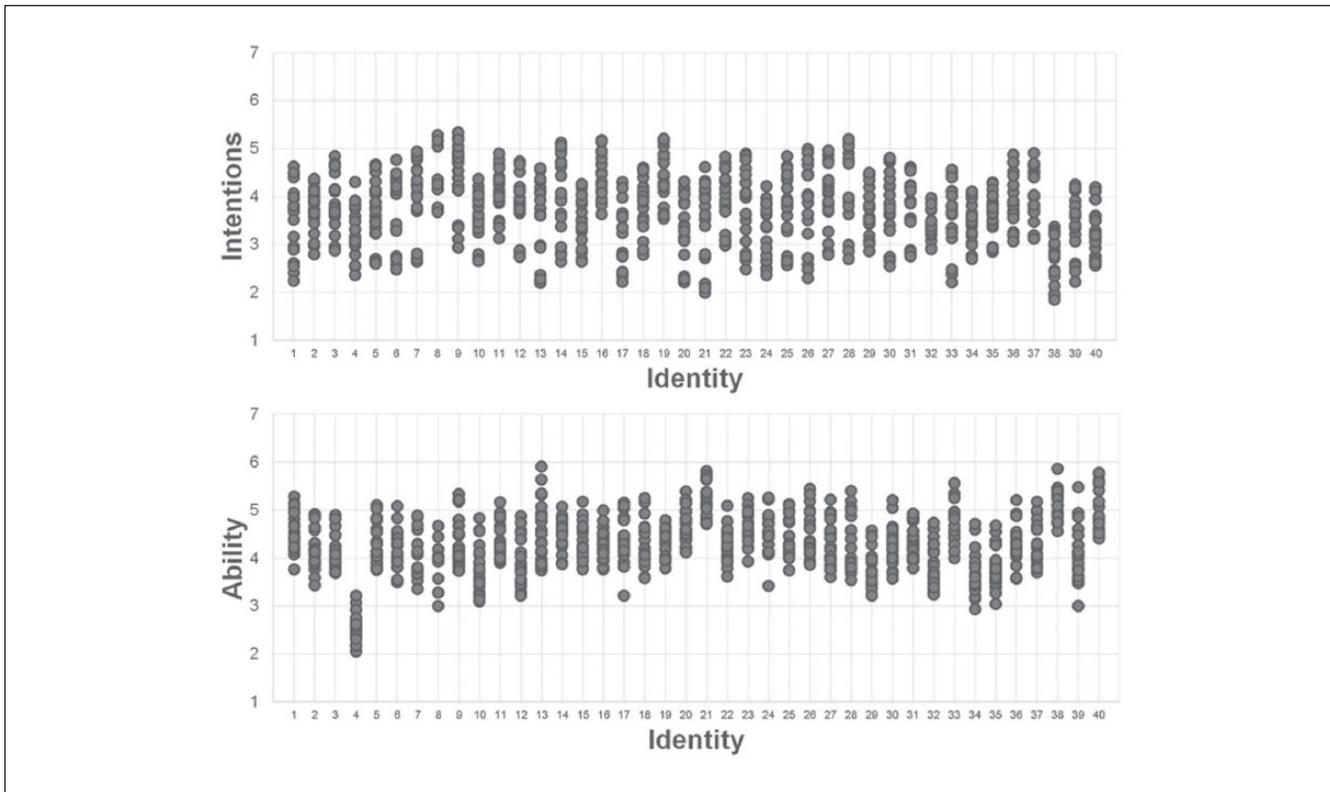


Figure 4. The ratings for each stimulus face across all forty identities in Study 2, demonstrating the greater variability in ratings of intentions than of ability.

them that, given the recent banking crisis in which many financial investors made risky and unscrupulous investment decisions, the purpose of the study was to examine how people select their financial advisors based on appearance. Previous research has revealed that when making decisions that could affect their economic resources, participants choose to interact with individuals they perceive to be good-intentioned and trustworthy (Kubota, Li, Bar-David, Banaji, & Phelps, 2013; Stanley, Sokol-Hessner, Banaji, & Phelps, 2011). This therefore provided a context in which the importance of perceived intentions would make subtle emotional resemblances potentially determinant of selection. Intentions or trustworthiness was never explicitly mentioned. In the Power-lifting condition, participants read a similar paragraph but instead were told the study was investigating sports betting in professional power lifting (see Appendix). This therefore provided a context in which the importance of perceived ability would make subtle differences in facial structure potentially determinant of selection. Physical ability was never explicitly mentioned. Participants were tasked with selecting four faces that were most likely to win their power-lifting competitions, a context in which the importance of perceived ability would make fWHR potentially determinant of selection. Previous research has shown that when making such decisions, participants tend to select targets with higher fWHR, who convey greater ability (Hehman, Leitner, & Freeman, 2014). Given prior

studies, we expected that faces conveying more positive intentions would be selected as financial advisors, whereas faces conveying more ability would be selected as power-lifting winners. Our critical hypothesis was that financial advisor selection would be more variable across multiple instances of a target (due to the malleability of perceived intentions), whereas power-lifting winner selection would be less variable (due to the stability of perceived ability).

Results

Manipulation check. We conducted a 2 (Emotion: Happy, Angry) \times 2 (fWHR: High, Low) \times 2 (Selection: Financial Advisor, Power-lifter) mixed-model ANOVA with repeated measures on the first two factors. In the Financial Advisor condition in which intentions should be of primary importance, the emotional resemblance of targets was most impactful in target selection. Participants selected targets with more positive facial resemblances, $F(1, 51) = 28.77, p < .00001, \eta_p^2 = .36$. In contrast, in the Power-lifter condition in which ability was expected to be primary, participants were most sensitive to the static facial structure of faces, being more likely to select targets with wider facial structure, $F(1, 50) = 48.24, p < .00001, \eta_p^2 = .49$.

As in Study 2, these dimensions were not completely independent. When selecting financial advisors, participants were

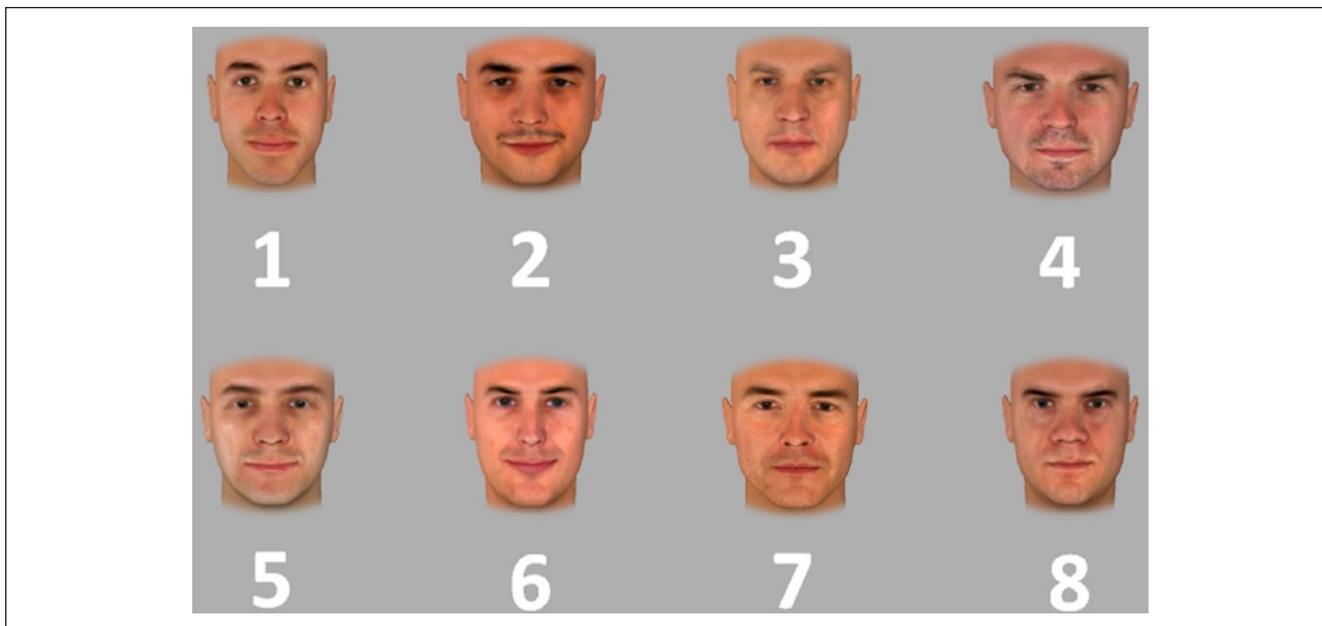


Figure 5. Example target array from Study 3.

also more likely to select targets with thinner facial structure, $F(1, 51) = 6.36, p = .015, \eta_p^2 = .11$. Furthermore, when selecting power-lifters, participants were additionally more likely to select more angry than happy faces, $F(1, 50) = 34.52, p < .00001, \eta_p^2 = .41$. The smaller relative effect sizes, however, indicate that these effects are secondary to our primary hypothesized effects.

Thus, in the Financial Advisor condition in which intentions were most important, participants' selections were more guided by dynamic emotional resemblance than static facial structure, whereas the reverse occurred in the Power-lifter condition in which ability was most important. Accordingly, these results show that emotional resemblance influences perceived intentions, which influences who one selects for a financial advisor. Similarly, facial structure influences perceived ability, which influences who one selects as a power-lifting winner. Thus, these basic evaluations predict important downstream social decisions.

Variance analysis. Testing our primary hypothesis regarding heterogeneity, we first examined whether there was greater variability in faces selected for intentions (emotional resemblance) than for ability (facial structure). To do so, we calculated the percentage of times faces with happy (vs. angry) emotional resemblance and wider (vs. thinner) facial structure were selected across all conditions. In contrast to Studies 1 and 2, heterogeneity within these percentages was assessed using Levene's test, as this design did not necessitate a multilevel framework. Results supported our hypothesis that there was greater variability in financial advisor decisions (based more on intentions and dynamic emotional resemblance) than power-lifting winner decisions (based

more on ability and static facial structure), although this effect was marginal, $F(1, 204) = 3.42, p = .066$.

Study 4

A limitation of the work to this point is that by measuring or manipulating stimuli along a priori dimensions, we have necessarily constrained the features that vary across the different faces. This may have reduced the contribution of other facial features potentially involved in evaluations of intentions and ability, and in turn, the selection of financial advisors and power-lifters. These facial manipulations and measurements are central to the results to this point, and thus, it is important to corroborate the findings of Studies 1 to 3 by assessing the perceptual basis of these various judgments without specifying any cues a priori. Accordingly, participants in the current study completed a reverse-correlation task, which allowed us to assess their representations of positive/negative intentions and low/high ability, as well as financial advisors and power-lifters. In addition, although the relationship between emotional resemblance and evaluations of intentions has been well-demonstrated in the literature (Engell et al., 2007; Oosterhof & Todorov, 2009), that facial structure was the primary driver of evaluations of ability, over and above emotional resemblance in Studies 2 and 3, is a novel premise. Accordingly, an additional goal of Study 4 was to further examine this possibility.

Method

Participants. Eighty-two (41 female, Age $M = 32$) participants completed the reverse-correlation task through Ama-

zon's Mechanical Turk for monetary compensation.

Stimuli and procedure. Reverse correlation is a data-driven approach, one variant of which involves generated random visual noise placed over a base-face. Over 100 trials, across which the apparent features of the base-face were randomly varied, participants selected 1 of 2 presented faces, and in this way the specific features predicting specific judgments were determined (Dotsch & Todorov, 2011; Dotsch, Wigboldus, Langner, & van Knippenberg, 2008). This data-driven technique is advantageous because no restrictions on which features might be diagnostic are implemented by the researchers. The base-face for the reverse-correlation task was created by morphing three male faces from a commercially available database of face images (www.3d.sk). The final image was gray-scaled and blurred. Following the reverse-correlation approach taken in other research (Dotsch & Todorov, 2011; Dotsch et al., 2008), participants were presented with two faces side-by-side on each trial. These faces were the same base-face over which randomly generated visual noise had been placed. Assessing the role of facial cues in explicit social evaluations, half the participants were randomly assigned to select the face that looked more trustworthy or higher in ability, the single ratings that best captured the dimensions of interest. The other half of participants read a paragraph similar to that in Study 3 about either financial advisors or power-lifters, and subsequently selected either the face they would prefer for their financial advisor or thought would be more likely to win a power-lifting competition.

Data preprocessing. Following the procedure of previous research (Dotsch & Todorov, 2011), for each participant we averaged the face selected on each trial, and then averaged the faces created by all participants to create four grand-average faces: Trustworthy and High Ability, and from the Financial Advisor and Power-lifter conditions.

Results

Visual inspection of the grand-average faces derived from explicit social evaluations (intentions and ability) were consistent with hypotheses: The grand-average face for trustworthiness had a happier emotional resemblance, whereas the grand-average face for ability was wider and had an angrier emotional resemblance (Figure 6).

Objective metrics confirmed this inspection. External raters ($n = 92$) were presented the grand-average faces from each of the four conditions. The Trustworthy and High-Ability faces were presented side-by-side, as were the faces from the Financial Advisor and Power-lifter conditions (which face appeared on the left vs. right was counterbalanced across participants). On a between-subjects basis, these participants rated which of the two faces in both pairs appeared more trustworthy or higher in ability. Paired-samples t tests further confirmed the utility of our earlier manipulations: 83% of the participants rated the

grand-average Trustworthy face as more trustworthy than the Physically Strong face, $t(70) = 7.39, p < .001, 95\% \text{ CI} = [.483, .841]$, and 64% chose the Physically Strong face as being stronger, $t(72) = 2.56, p = .013, 95\% \text{ CI} = [.063, .513]$.

Similarly, the grand-average face from the Financial Advisor condition, a context in which intentions is important, was rated as more trustworthy than the face from the Power-lifter condition. Eighty-seven percent of the participants rated this face as more trustworthy, $t(70) = 9.39, p < .001, 95\% \text{ CI} = [0.588, 0.905]$. Seventy-three percent rated the Power-lifter face, created in a context in which ability was important, as physically stronger, $t(72) = 4.30, p < .001, 95\% \text{ CI} = [.243, .662]$.

These results corroborate the findings of Studies 1 to 3 using a reverse-correlation paradigm that makes no a priori assumptions about the cues relevant for these social evaluations. The findings provide further support that emotional resemblance primarily contributes to evaluations of intentions, in turn driving decisions about financial advisors, a process that is malleable and varies across multiple instances of a target. Facial structure, on the other hand, primarily contributes to evaluations of ability, in turn driving decisions about power-lifters, a process that is more stable and consistent across multiple instances of a target.

General Discussion

Faces are a strong influence on social evaluation, and thus, it is not surprising they have received much attention over the past several decades (Zebrowitz, 2006). A great deal of research has explored the issue of consensus in social evaluation, examining perceiver consensus with the target (Moskowitz, 1990), across multiple raters (Berry, 1991), and the factors that moderate the degree of consensus (Ambady & Gray, 2002; Kenny & Albright, 1987; Kenny, 1991). Only surprisingly recently has research explored how consensus plays out across multiple instances of a single target (Jenkins, White, Van Montfort, & Burton, 2011; Murphy et al., 2015; Todorov & Porter, 2014), revealing the significant instability of some social evaluations.

The current research provides insight into when social evaluations across multiple instances of a target are more or less likely to be consistent. In Kenny's (1991) Consensus–Accuracy model, one parameter critical to consensus is consistency, or the stability of a target's appearance or behavior across multiple presentations. The current work demonstrates that because of their underlying physiological basis (i.e., bone structure vs. facial musculature), different types of appearance cues (i.e., static vs. dynamic) have a different degree of consistency. Cues originating in more dynamic features (e.g., emotional resemblance) vary to a greater extent across multiple presentations of a single target, and thus, evaluations based on these cues, such as intentions, were less consistent. Evaluations of intentions were malleable and fluctuated with momentary emotional resemblances. In contrast, we found that cues originating from more static

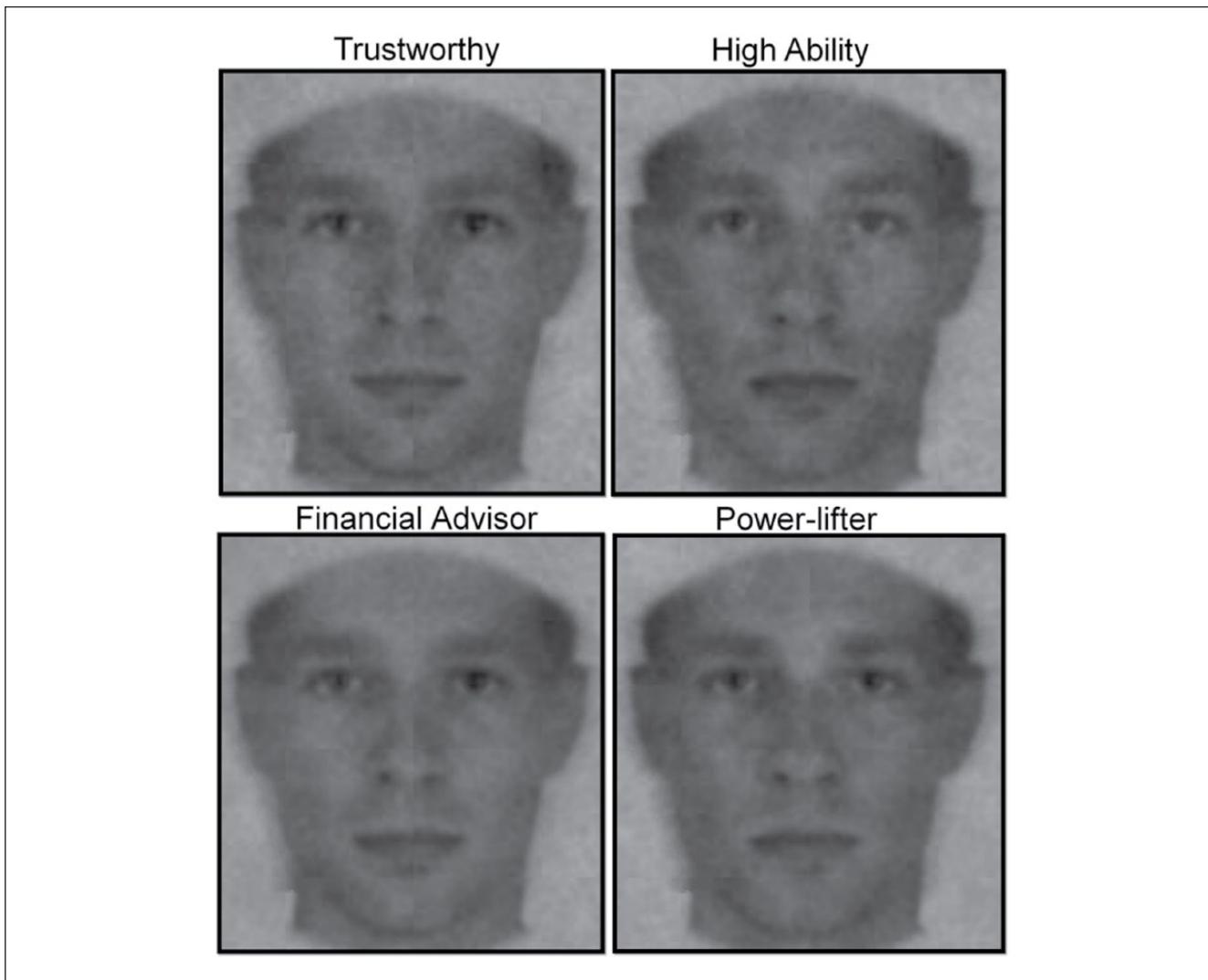


Figure 6. Grand average faces from all conditions.

features (e.g., facial structure) were more consistent across multiple presentations, and thus, evaluations based on these cues, such as ability, were more consistent. Furthermore, the consistency in these evaluations had downstream implications. Greater consistency in perceived ability led perceivers’ decisions based on this evaluation to be relatively stable across multiple photos of a target. In contrast, lesser consistency in perceived intentions led perceivers’ decisions based on intentions to be relatively malleable across multiple instances of a target.

Thus, appearance cues vary upon a static to dynamic continuum, and we demonstrate that social evaluations arising from these different types of cues will have differing levels of consistency. It is important to stress, however, that rather than discrete categories of “dynamic” and “static” features, we believe it is best to conceptualize each feature upon this continuum. Although muscles are to some extent fluid, they are attached to the underlying bone structure, and so some

degree of emotional resemblance may come from facial structure, and vice versa. Demonstrating that even relatively static features can be more dynamic in some behaviors and contexts, individuals spontaneously manipulated fWHR by tilting their heads to appear more intimidating (Hehman, Leitner, & Gaertner, 2013). Researchers might be mindful that the relative static versus dynamic nature of each cue might further vary depending on the stimuli, context, and behavior being examined.

Although we have focused on two social dimensions, intentions and ability, we speculate any number of dimensions would be candidates for the effects observed here. For instance, though we focused on how perceived ability arises from static facial bone structure, other static cues are likely to give rise to unique but equally important social evaluations. There are numerous facial cues utilized in social evaluation, and these naturally vary in their relatively dynamic versus static nature. Future work could more comprehensively test a wider gamut

of facial cues in driving relatively malleable or stable evaluations across multiple instances of a social target. For now, however, the current research creates an important distinction between relatively dynamic and static facial features and their influences on the consistency of social evaluations.

To this point, we have refrained from discussing the broad literature examining accuracy, or perceivers making inferences from target appearances that are to some degree associated with behavior or self-reports. Consensus is neither a necessary nor sufficient condition for accuracy, although normally the two are closely tied (Kenny, 1991), and our results here are agnostic with regard to accuracy. However, one implication of our results is that accurate evaluations may more readily arise from static rather than dynamic appearance cues. Due to lesser variability in static than dynamic cues across multiple presentations of a target, there would be decreased error in social evaluations, and any honest signal would have a stronger opportunity to avail itself. That said, the results raise an intriguing question regarding the nature of dynamic cues' variability across multiple presentations. For example, dynamic emotional resemblances may indeed vary across multiple instances of a target, yet in specific contexts exhibit a chronic tendency that conveys accurate information. For instance, although smiling is a dynamic cue, the frequency of a subtle smile might be diagnostic of intentions. If true, it implies that evaluations based on dynamic cues also have the potential for accuracy, but accuracy might be more likely to emerge in specific contexts where a stable tendency might exist. Thus, rather than attempt to control for factors such as emotional resemblance or self-presentation in accuracy studies, one could instead conceive of such variability as meaningful. In contrast, for evaluations supported primarily by static cues, whether the signal conveyed is accurate or error-prone might be far less context-specific. Thus, although our results cannot directly speak to questions of accuracy, they suggest that certain evaluations may have greater opportunities for accuracy than others and highlight the possible importance of social context.

Finally, our findings additionally have important methodological implications. The variability of social evaluations across different images of a single person may introduce problems when studying social evaluation using only single images of a target, a widespread and commonplace approach. Together with recent work (Jenkins et al., 2011; Rule et al., 2013; Todorov & Porter, 2014), these findings highlight an important question facing a great deal of person perception research as to whether evaluations of targets reflect stable inferences of the target versus inferences that are highly susceptible to momentary changes across the target's photos due to dynamic features. When correlating evaluations supported by static cues with external variables (e.g., behavior, objective outcomes, personality), significant relationships might emerge with a fewer number of ratings or participants than

evaluations supported by dynamic cues. Thus, understanding the relatively static versus dynamic nature of the facial cues supporting the evaluations in question is critical. Furthermore, assessing relationships between evaluations supported by dynamic cues with external variables might require larger samples, greater statistical power, or different methodological approaches. Different approaches, such as presenting additional or more representative images of targets, and methods to assess whether a presentation of a target is representative of that target's behavior, would help to circumvent the issues presented here.

In conclusion, we found that social evaluations arising from dynamic features vary to a greater extent across multiple instances of a single target than those based on static features. Specifically, we found that perceived intentions of a target are malleable and contingent on more moment-to-moment changes in emotion resemblance, whereas perceived ability is more fixed and tied to the face's static structure. Our findings highlight the need to appreciate the consistency in the facial cues supporting certain evaluations to understand the consistency of those evaluations themselves.

Authors' Note

E.H. and J.B.F. designed the experiments. E.H. collected the data. E.H. and J.K.F. analyzed the data. All authors contributed to writing the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. This test of variance was conducted using 10 Level 2 clusters. Simulation studies have indicated that in multilevel frameworks, estimates of variance can sometimes be biased with a smaller number of Level 2 clusters (Maas & Hox, 2005). Due to this concern, we additionally calculated the amount of variance in each rating outside this multilevel framework. Results were identical.
2. As we are interested in evaluations of emotionally neutral faces, emotional resemblance was purposely made extremely subtle such that if perceived alone each face might be evaluated as neutral.

Supplemental Material

The online supplemental material is available at <http://pspb.sagepub.com/supplemental>.

References

- Ambady, N., Bernieri, F., & Richeson, J. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices

- of the behavioral stream. *Advances in Experimental Social Psychology*, 32, 201-271.
- Ambady, N., & Gray, H. M. (2002). On being sad and mistaken: Mood effects on the accuracy of thin-slice judgments. *Journal of Personality and Social Psychology*, 83, 947-961. doi:10.1037//0022-3514.83.4.947
- Berry, D. S. (1991). Accuracy in social perception: Contributions of facial and vocal information. *Journal of Personality and Social Psychology*, 61, 298-307. doi:10.1037//0022-3514.61.2.298
- Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques—SIGGRAPH*, 99, 187-194. doi:10.1145/311535.311556
- Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer (Eds.), *Advances in social cognition* (pp. 1-36). Hillsdale, NJ: Lawrence Erlbaum.
- Carré, J. M., & McCormick, C. M. (2008). In your face: Facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proceedings of the Royal Society B: Biological Sciences*, 275, 2651-2656. doi:10.1098/rspb.2008.0873
- Carré, J. M., McCormick, C. M., & Mondloch, C. J. (2009). Facial structure is a reliable cue of aggressive behavior. *Psychological Science*, 20, 1194-1198. doi:10.1111/j.1467-9280.2009.02423.x
- Carré, J. M., Morrissey, M. D., Mondloch, C. J., & McCormick, C. M. (2010). Estimating aggression from emotionally neutral faces: Which facial cues are diagnostic? *Perception*, 39, 356-377. doi:10.1068/p6543
- Darwin, C. (1872). *The expression of the emotions in man and animals*. London, England: Murray.
- Dotsch, R., & Todorov, A. (2011). Reverse correlating social face perception. *Social Psychological & Personality Science*, 3, 562-571. doi:10.1177/1948550611430272
- Dotsch, R., Wigboldus, D. H. J., Langner, O., & van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, 19, 978-980. doi:10.1111/j.1467-9280.2008.02186.x
- Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience*, 19, 1508-1519. doi:10.1162/jocn.2007.19.9.1508
- Erikson, E. H. (1950). *Childhood and society*. New York, NY: W.W. Norton.
- Farkas, L. G. (1994). *Anthropometry of the head and face*. New York, NY: Raven Press.
- Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology*, 23, 1-74.
- Freedman, M. B., Leary, T. F., Ossario, A. G., & Coffey, H. S. (1953). The interpersonal dimension of personality. *Journal of Personality*, 20, 143-161.
- Hehman, E., Carpinella, C. M., Johnson, K. L., Leitner, J. B., & Freeman, J. B. (2014). Early processing of gendered facial cues predicts the electoral success of female politicians. *Social Psychological & Personality Science*, 5, 815-824. doi:10.1177/1948550614534701
- Hehman, E., Leitner, J. B., Deegan, M. P., & Gaertner, S. L. (2013). Facial structure is indicative of explicit support for prejudicial beliefs. *Psychological Science*, 24, 289-296. doi:10.1177/0956797612451467
- Hehman, E., Leitner, J. B., Deegan, M. P., & Gaertner, S. L. (2015). Picking teams: When dominant facial structure is preferred. *Journal of Experimental Social Psychology*, 59, 51-59. doi:10.1016/j.jesp.2015.03.007
- Hehman, E., Leitner, J. B., & Freeman, J. B. (2014). The face-time continuum: Lifespan changes in facial width-to-height ratio impact aging-associated perceptions. *Personality and Social Psychology Bulletin*, 40, 1624-1636. doi:10.1177/0146167214552791
- Hehman, E., Leitner, J. B., & Gaertner, S. L. (2013). Enhancing static facial features increases intimidation. *Journal of Experimental Social Psychology*, 49, 747-754. doi:10.1016/j.jesp.2013.02.015
- Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, 121, 313-323. doi:10.1016/j.cognition.2011.08.001
- Kenny, D. A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review*, 98, 155-163. doi:10.1037//0033-295X.98.2.155
- Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin*, 102, 390-402.
- Kubota, J. T., Li, J., Bar-David, E., Banaji, M. R., & Phelps, E. A. (2013). The price of racial bias: Intergroup negotiations in the ultimatum game. *Psychological Science*, 24, 2498-2504. doi:10.1177/0956797613496435
- Lefevre, C. E., Lewis, G. J., Perrett, D. I., & Penke, L. (2013). Telling facial metrics: Facial width is associated with testosterone levels in men. *Evolution & Human Behavior*, 34, 273-279. doi:10.1016/j.evolhumbehav.2013.03.005
- Maas, C., & Hox, J. (2005). Sufficient sample sizes for multi-level modeling. *Methodology*, 1, 86-92. doi:10.1027/1614-1881.1.3.86
- McDonald, M. M., Asher, B. D., Kerr, N. L., & Navarrete, C. D. (2011). Fertility and intergroup bias in racial and minimal-group contexts: Evidence for shared architecture. *Psychological Science*, 22, 860-865. doi:10.1177/0956797611410985
- Moskowitz, D. S. (1990). Convergence of self-reports and independent observers: Dominance and friendliness. *Journal of Personality and Social Psychology*, 58, 1096-1106. doi:10.1037//0022-3514.58.6.1096
- Murphy, N. A., Hall, J. A., Schmid Mast, M., Ruben, M. A., Frauendorfer, D., Blanch-Hartigan, D., . . . Nguyen, L. (2015). Reliability and validity of nonverbal thin slices in social interactions. *Personality and Social Psychology Bulletin*, 41, 199-213. doi:10.1177/0146167214559902
- Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 11087-11092. doi:10.1073/pnas.0805664105
- Oosterhof, N. N., & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion*, 9, 128-133. doi:10.1037/a0014520
- Phillips, P., Wechsler, H., Huang, J., & Rauss, P. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16, 295-306.

- Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology, 104*, 409-426. doi:10.1037/a0031050
- Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 7710-7715. doi:10.1073/pnas.1014345108
- Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*, 1623-1626. doi: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*, 1404-1417. doi:10.1177/0956797614532474
- Tooby, J., & Cosmides, L. (1988, April). *The evolution of war and its cognitive foundations* (Institute for Evolutionary Studies Technical Report 88-1). Retrieved from https://www.academia.edu/4662770/The_Evolution_of_War_and_its_Cognitive_Foundations
- Van Vugt, M. (2009). Sex differences in intergroup competition, aggression, and warfare: The male warrior hypothesis. *Annals of the New York Academy of Sciences, 1167*, 124-134. doi:10.1111/j.1749-6632.2009.04539.x
- Wong, E. M., Ormiston, M. E., & Haselhuhn, M. P. (2011). A face only an investor could love: CEOs' facial structure predicts their firms' financial performance. *Psychological Science, 22*, 1478-1483. doi:10.1177/0956797611418838
- Zebrowitz, L. A. (2006). Finally, faces find favor. *Social Cognition, 24*, 657-701. doi:10.1521/soco.2006.24.5.657
- Zebrowitz, L. A., Androletti, C., Collins, M. A., Lee, S. Y., & Blumenthal, J. (1998). Bright, bad, babyfaced boys: Appearance stereotypes do not always yield self-fulfilling prophecy effects. *Journal of Personality and Social Psychology, 75*, 1300-1320.
- Zebrowitz, L. A., Fellous, J.-M., Mignault, A., & Androletti, C. (2003). Trait impressions as overgeneralized responses to adaptively significant facial qualities: Evidence from connectionist modeling. *Personality and Social Psychology Review, 7*(3), 194-215.
- Zebrowitz, L. A., & McDonald, S. (1991). The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior, 15*, 603-623.