Individual differences in trust evaluations are shaped mostly by environments, not genes

Clare A. M. Sutherland[a,b,1], Nichola S. Burton[a], Jeremy B. Wilmer[c], Gabriëlla A. M. Blokland[d,e], Laura Germine[f,g], Romina Palermo[a], Jemma R. Collova[a], and Gillian Rhodes[a]

[a]Australian Research Council Centre of Excellence in Cognition and its Disorders, School of Psychological Science, University of Western Australia, Crawley, WA 6009, Australia; [b]School of Psychology, King's College, University of Aberdeen, Aberdeen, AB24 3FX, Scotland; [c]Department of Psychology, Wellesley College, Wellesley, MA 02481; [d]Department of Psychiatry and Neuropsychology, Maastricht University, 6229 ER Maastricht, The Netherlands; [e]School for Mental Health and Neuroscience, Faculty of Health, Medicine, and Life Sciences, Maastricht University, 6229 ER Maastricht, The Netherlands; [f]McLean Institute for Technology in Psychiatry, McLean Hospital, Belmont, MA 02478; and [g]Department of Psychiatry, Harvard Medical School, Boston, MA 02115

Edited by Marlene Behrmann, Carnegie Mellon University, Pittsburgh, PA, and approved March 11, 2020 (received for review November 15, 2019)

People evaluate a stranger's trustworthiness from their facial features in a fraction of a second, despite common advice "not to judge a book by its cover." Evaluations of trustworthiness have critical and widespread social impact, predicting financial lending, mate selection, and even criminal justice outcomes. Consequently, understanding how people perceive trustworthiness from faces has been a major focus of scientific inquiry, and detailed models explain how consensus impressions of trustworthiness are driven by facial attributes. However, facial impression models do not consider variation between observers. Here, we develop a sensitive test of trustworthiness evaluation and use it to document substantial, stable individual differences in trustworthiness impressions. Via a twin study, we show that these individual differences are largely shaped by variation in personal experience, rather than genes or shared environments. Finally, using multivariate twin modeling, we show that variation in trustworthiness evaluation is specific, dissociating from other key facial evaluations of dominance and attractiveness. Our finding that variation in facial trustworthiness evaluation is driven mostly by personal experience represents a rare example of a core social perceptual capacity being predominantly shaped by a person's unique environment. Notably, it stands in sharp contrast to variation in facial recognition ability, which is driven mostly by genes. Our study provides insights into the development of the social brain, offers a different perspective on disagreement in trust in wider society, and motivates new research into the origins and potential malleability of face evaluation, a critical aspect of human social cognition.

trust | face evaluation | first impressions | behavioral genetics | classical twin design

Trust is fundamental to human life: Without trust, society itself would not exist (1, 2). Who we trust and why have become defining questions of our era as public trust in expertise is rapidly declining in favor of trust based on rather more superficial characteristics (2). For example, although popular advice compels us not to "judge a book by its cover," people readily place their trust in others based on a superficial evaluation of their facial features. Trust evaluations from faces wield widespread social influence, predicting financial lending, consumer choices, and criminal justice outcomes (3). At the extreme, untrustworthy-looking people are more likely to receive the death penalty than trustworthy-looking people, when convicted of a major crime (4). Influential theories explain how face evaluations are formed, including identifying trustworthiness as the primary dimension of face evaluation (5–7) and suggesting that trustworthiness evaluation is adaptive because it functions to detect threat (5–7). Unfortunately, whereas trust decisions are fundamentally formed on an individual level, the science of facial trust has mostly operated by examining consensus trust evaluations at a group level (e.g., this face looks more trustworthy than another to most people). Although certain facial characteristics are broadly considered by most people to look trustworthy (e.g., smiling, femininity, head tilt) (7, 8), very recent research has suggested that people may disagree to a surprising extent when judging trustworthiness from faces (9–11). Ignoring individual trust evaluation is problematic for two reasons: first, because current theory becomes disconnected from everyday, individual decision making and, second, because we lack directly meaningful data as to the degree of plasticity or capacity for change in facial trust evaluation.

Here, we set out to provide an unprecedentedly thorough investigation of the nature of individual trust impressions and the degree to which they are shaped by genetics, shared life experience, or personal life experience. We started by developing reliable tests of individual differences in facial trust evaluation, using best-practice iterative test development procedures. We used these tests to confirm that trustworthiness impressions are, to a substantial degree, "in the eye of the beholder." Despite the overwhelming focus of previous work on consensus impressions, therefore, trustworthiness impressions are also based on stable individual differences that are unique to the individual. This

Significance

Rapid impressions of trustworthiness can have extreme consequences, impacting financial lending, partner selection, and death-penalty sentencing decisions. But to what extent do people disagree about who looks trustworthy, and why? Here, we demonstrate that individual differences in trustworthiness and other impressions are substantial and stable, agreeing with the classic idea that social perception can be influenced in part by the "eye of the beholder." Moreover, by examining twins, we show that individual differences in impressions of trustworthiness are shaped mostly by personal experiences, instead of genes or familial experiences. Our study highlights individual social learning as a key mechanism by which we individually come to trust others, with potentially profound consequences for everyday trust decisions.

Author contributions: C.A.M.S., N.S.B., J.B.W., G.A.M.B., L.G., R.P., and G.R. designed research; C.A.M.S., N.S.B., J.B.W., and J.R.C. performed research; C.A.M.S., N.S.B., J.B.W., and G.A.M.B. contributed new reagents/analytic tools; C.A.M.S. and N.S.B. analyzed data; C.A.M.S., N.S.B., J.B.W., and G.R. wrote the paper; C.A.M.S., J.B.W., and G.R. planned and supervised the research; C.A.M.S., N.S.B., and J.R.C. managed twin recruitment; and G.A.M.B., L.G., R.P., and J.R.C. helped shape the interpretation and provided critical feedback on the manuscript draft.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Data deposition: Data, code, and materials are available at the Open Science Framework, https://osf.io/35zf8/?view_only=e76c6755dcea4be2adc5b075cae896e8. The face impression tests can be viewed at https://www.testable.org/experiment/855/674205/start.

[1]To whom correspondence may be addressed. Email: clare.sutherland@abdn.ac.uk.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1920131117/-/DCSupplemental.

First published April 27, 2020.

disagreement in trust evaluation of faces parallels a more widespread general trend in society as increasingly stark disagreement is observed in whom we consider to be trustworthy (1).

The existence of large and stable individual differences in facial trustworthiness evaluation in turn raises the key question of why people vary in their trustworthiness impressions. That is, what casual mechanism(s) explain these individual differences? Core face-processing abilities, such as individual differences in face identity recognition, have been shown to be almost entirely driven by genes (12–14). Trusting personality traits also show strong genetic influences (15) as does general social trust (16). Genetic factors might therefore also shape variation in facial trustworthiness evaluation, especially given that trustworthiness impressions are based on facial features that overlap with those used for identity recognition (for example, eyebrow height, face width, and so on) (7, 17). Indeed, major genetic influences on human behavior are so ubiquitous that, to be complete, social psychological theories of traits must consider such genetic influences (15). Yet, reliably measured traits that are primarily shaped by environment factors, although rare in behavioral genetics, do exist. Strikingly, for example, individual differences in facial attractiveness evaluations result primarily from individual experiences rather than genes (18), and attractive and trustworthy facial features partially overlap (for example, smiling is both attractive and trustworthy) (19). Moreover, trustworthiness evaluation has recently been shown to be shaped by contextual factors, such that more typical (20), more familiar (21), and more common (22) faces look more trustworthy. Experiences—in particular, the "diet" of faces one encounters in daily life—may therefore play an important role in trustworthiness perception. Two individuals who experience different facial variation in their environments, whether in real life or online, may learn to trust different facial features (for example, blue or green eyes, narrow or wide facial structure, and so on) as these particular features become familiar. Social learning has also recently been shown to influence trustworthiness evaluation: For example, people are more likely to trust others who resemble trustworthy individuals encountered in an initial economic transaction (23). Individual trust interactions with others may therefore also act to drive variation in trust evaluation between people. For example, after experiencing trustworthy behavior from feminine-looking individuals, one learns to trust feminine features in particular. These theories instead predict that either shared or personal environments are especially critical for variation in trustworthiness evaluation, rather than genes. Thus, previous studies on face recognition and general social trust predict a genetic basis to face trustworthiness evaluation whereas social and statistical learning theories predict a shared or personal environmental contribution instead.

Here, we carried out a twin study, allowing us to measure the genetic and environmental contributions to individual differences in face trustworthiness evaluation. We found that variation in face trustworthiness evaluation is largely shaped by variation in personal environments, rather than genes or shared familial environment. Our findings represent a rare example of a core social perceptual capacity being predominantly shaped by a person's personal environment. We also investigated the specificity of individual differences in trustworthiness evaluation, finding that variation in trustworthiness impressions dissociates from variation in other core social perceptual judgments of attractiveness and dominance. Using multivariate twin modeling, we demonstrate that different personal experiences drive variation in each trait. Finally, we show that individual differences in impressions also dissociate from identity and emotion recognition, representing core abilities in other major face perception theories (24). Our findings provide insight into the development of the social brain by demonstrating that a distinctive etiological architecture underlies different aspects of variation in face perception. Understanding variation in face perception has recently become a major new direction of interest across scientific fields, from visual science (18, 25) and social cognition (9) to psychometrics and cognitive neuroscience (23).

Our aim to understand to what extent and why people vary in face trustworthiness evaluation is timely: We are currently facing a global societal shift in whom we trust, as trust in institutions and objective expertise is replaced by trust in individuals and personal connections (1, 2). With the rise of the internet, social worlds have fragmented into echo chambers while political and cultural divides widen (2). These echo chambers create potent individual microenvironments of trust and mistrust. Simultaneously, disagreement on who is trustworthy has never been higher. Our findings offer a different perspective to understand to what extent and why we disagree when deciding whom to trust: Exposure to different social information is itself a driver of disagreement. Moreover, as the information we access online becomes increasingly individual, our findings also suggest that differences in whom we trust will further widen. Since individual differences in face evaluation are primarily learned through experience, then these impressions may also be more malleable than previously suggested (17). Modifying face trustworthiness evaluation has considerable practical importance, given that impressions of trustworthiness can bias real social and economic outcomes (3).

Results

Our primary focus was on trustworthiness, given the central importance of trust to human society. In order to understand the specificity of our results with regard to trustworthiness, we additionally included dominance and attractiveness control tasks as these judgments represent other major dimensions of social evaluation of faces (7, 8) and also wield considerable influence on everyday decisions, including promotion choices (3) and partner selection (26). We first developed standardized tests that were capable of distinguishing stable individual differences in trustworthiness, dominance, and attractiveness evaluation from random or inconsistent responding. We developed such tests by initially generating a large pool of candidate face stimuli, selected for ecological validity, then iteratively refining reliability and construct validity over three progressive waves of data collection (total $n = 1{,}344$) (*Materials and Methods* and *SI Appendix*, *Test development and reliability*). Tests are available for future research. We then used the refined tests to measure face impressions in a large sample of 1,264 twin individuals (*Materials and Methods* and *SI Appendix*, Table S3).

Despite the overwhelming focus in previous work on consensus impressions of faces (reviewed in ref. 17), we found that a substantial, stable, and significant proportion of variance in trustworthiness, dominance, and attractiveness evaluations were due to unique face preferences that differed across observers (Fig. 1 and *SI Appendix*, Fig. S1).

Individual differences in facial impressions were quantified using a multilevel modeling intraclass correlation method that parsed variation in impressions. Specifically, we fit an intercept-only multilevel model with random effects to estimate intraclass correlations (ICCs) that reflect the proportion of variance in ratings that can be explained by the faces, the participants, and the interaction between faces and participants (*SI Appendix*, *Measuring individual differences in facial impressions*). Critically, individual differences in facial impressions are captured in the model by the interaction between faces and participants, representing unique face preferences that differ across participants (9, 27). For example, one participant may consistently view narrow faces as more trustworthy than wide faces whereas another participant has the opposite perception. Consensus face impressions and individual differences not directly related to faces, such as overall trust, scale use, and so on, were captured in the model as main effects for faces and participants (*SI Appendix*,

PSYCHOLOGICAL AND COGNITIVE SCIENCES

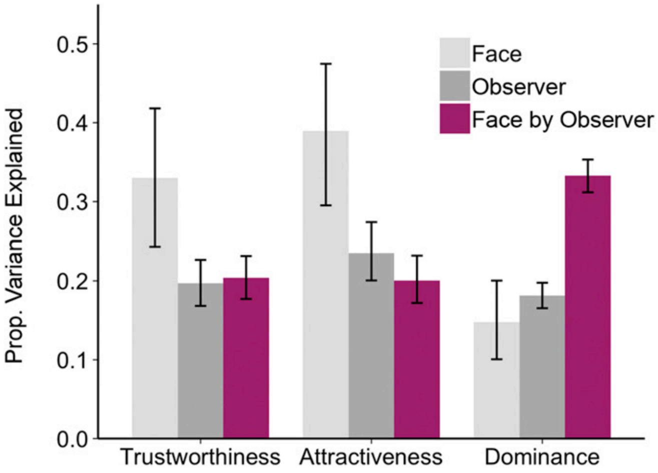

Fig. 1. Quantifying and explaining individual variation in face evaluation. Variance in key facial impressions explained by Face, Observer, and the unique combination of Face by Observer that reflects individual differences in facial evaluation. Variance is computed through intraclass correlation coefficients (ICCs) in a random-intercept multilevel model. Data from twin sample. *N* observations = 379,200 (1,264 perceivers by three traits by 50 faces by two trials). Error bars are 95% confidence intervals, fit using bootMer from lme4 in R (2,000 bootstrap samples). Individual difference components are substantial and significant.

Analysis of mean impressions). The finding of substantial stable observer variation supports the classic idea that impressions can be considerably idiosyncratic, lying at least partially in the "eye of the beholder." This result also motivates our main question of where these individual differences come from. That is, what are the genetic or environmental origins of individual differences in facial evaluation of trustworthiness and other key traits?

We determined the origins of individual differences in facial impressions using a classical twin design. Specifically, we compared same-sex dizygotic (DZ) twins (157 twin pairs, 372 total individuals) and monozygotic (MZ) twins (333 twin pairs, 781 total individuals) (*Materials and Methods* and *SI Appendix*, Table S3). In order that similarity could be compared across DZ and MZ twin pairs, we calculated face impression scores for each twin to index individual differences in facial impressions (i.e., the extent to which they agree or not with consensus facial impressions of trustworthiness, dominance, or attractiveness) (Fig. 2*A*).

We found that variation in face trustworthiness evaluation was largely shaped by people's personal experiences, rather than genetic differences or shared environments (Fig. 2*A*). To measure the heritability of face trustworthiness evaluation, we built standard ACE twin models using OpenMx (version 2.11.5) in R (version 3.5) (28). The ACE model is a multigroup structural equation model which estimates the similarity between MZ and DZ twin pairs, respectively. Although MZ and DZ twins share family environment to a similar extent (29), MZ twins share, on average, twice as much genetic variation as DZ twins. Thus, additive genetic covariances are modeled differently for MZ twins (fixed to 100% as they have 100% of their genetic influences in common) and DZ twins (fixed to 50% as they have, on average, 50% of their genetic influences in common). There are three main pathways in the model: the A pathway, representing additive genetic covariance between twin pairs; the C pathway, representing shared environmental covariance between twin pairs; and the E pathway, representing unique or personal environmental influence and measurement error. A, C, and E estimates are derived by comparing MZ and DZ twin pair similarity against the obtained data, using maximum likelihood to determine the combination of parameters that best fits the observed data. We obtained 95% CIs for parameter estimates using likelihood-based confidence interval estimation in OpenMx. We obtained these confidence intervals so that each estimate can be statistically compared to zero to determine significance at $P < 0.05$.

In the ACE models, the contribution of personal environmental factors was 70 to 82% across all three facial impressions (Table 1 and Fig. 2*A*). For all impressions, additive genetic factors made the next-largest contribution although the genetic influence was far smaller and not statistically significant (17 to 30%) (Table 1 and Fig. 2*A*). Shared environmental factors made the smallest contribution, also not statistically significant (0 to 11%) (Table 1 and Fig. 2*A*). Systematic testing of simpler AE, CE, and E models found that the AE model (i.e., C == 0) showed the best fit for all three impressions; personal environmental factors explained the majority of the variance in all models tested.

Any estimation of the contribution of the personal environment includes measurement error; therefore, it is important to demonstrate good reliability of measurement in twin studies (29). Internal reliability for face impression scores was excellent ($r = 0.75$–0.82, reliability is split-half and corrected for attenuation; Table 1), and test–retest reliability was also acceptable ($r = 0.52$–0.73) as measured in an independent sample ($n = 94$) (*SI Appendix*, *Test development and reliability*). Critically, reliability was comparable to gold-standard tests in face perception (e.g., facial identity recognition) (30), especially for trustworthiness (Fig. 2*B*). Reliability was also highly similar to reliability of personality measures used in previous twin research (see ref. 15 for a review). Overall, the twins' impressions, collected online, also correlated highly with impressions from nontwins tested in the laboratory ($n = 214$) and online ($n = 94$, $r > 0.93$ for all tests) (*SI Appendix*, Table S1), demonstrating comparability between unsupervised online versus laboratory-based assessments, as well as between our twin sample and nontwin participants. Given the high reliability of our tests, even when measured conservatively (i.e., with an alternative-forms, test–retest procedure), the contribution of personal environmental factors to variation in facial impressions cannot be explained by unreliable measurement. Instead, our findings lead us to conclude that variation in facial impressions is primarily driven by unique life experiences.

Results were unchanged after controlling for performance on a control (scene) evaluation task, age, sex, their products and powers (second and third), and twin birth order (*SI Appendix*, Table S6). Results were also unchanged if unpaired twins were excluded, indicating that attrition in our twin sample did not affect conclusions. Assumption tests for twin modeling were met, including homogeneity across twin pairs and zygosity for both means and variances (models corrected for outliers) (*SI Appendix*, *Twin analysis of individual differences in impressions*) (results did not change with outliers included).

Finally, the ACE modeling conclusions agreed with simple calculations based on comparing intraclass correlations between face impression scores for MZ twins with those for DZ twins (Fig. 3*A*). If the face impression scores of MZ twins are (significantly) more highly correlated than those of DZ twins, then there is a genetic contribution to individual differences in impressions. However, across all trait impressions, the 95% CIs for the MZ and DZ twin correlations entirely overlapped. The same conclusions were drawn when estimates were based on Falconer's formula (31) (Table 1 and *SI Appendix*, Table S4) and when maximum likelihood correlations were used (*SI Appendix*, Table S5). Across all of the tests, the majority of variance in individual differences in trustworthiness, dominance, and attractiveness evaluations was therefore attributable to nongenetic, environmental factors.

The strong personal environmental contribution to variation in facial evaluations found here stands in sharp contrast to

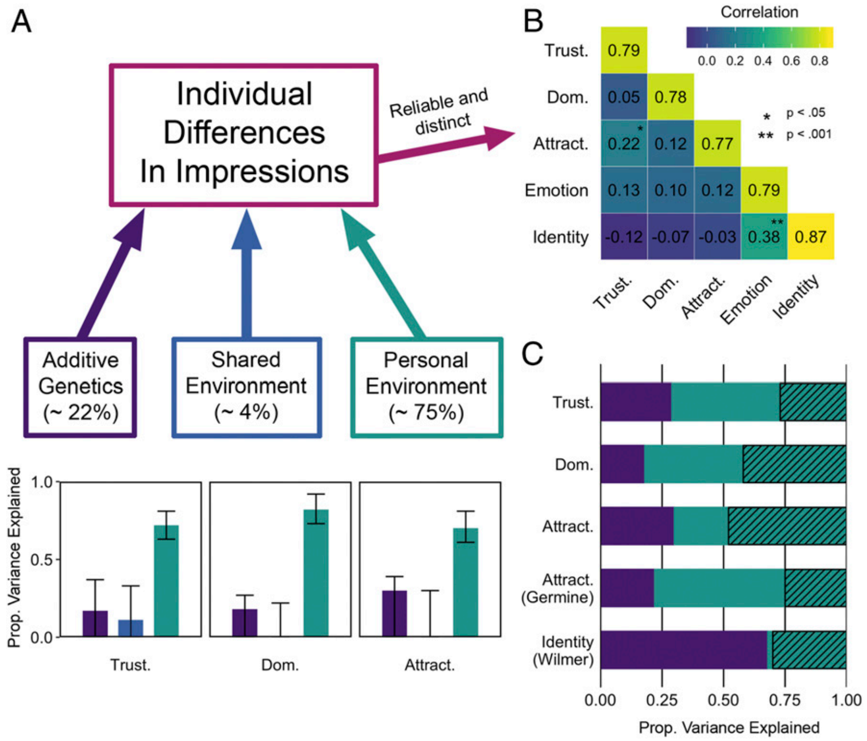

Fig. 2. (*A*) Additive genetic (purple), shared environment (blue), and personal environment contributions (green; including measurement error) to proportion (Prop.) variance explained in individual facial impressions, including ACE estimates averaged across the three impressions (boxes above) as well as individual univariate ACE models (bars below; n = 1,153, N observations = 3,426) (*SI Appendix*). Error bars represent 95% CIs with a lower bound of zero. (*B*) Individual differences in facial impressions are stable, with high internal reliability (shown on the diagonal, reliabilities are split-half and corrected for attenuation). Individual differences in facial impressions of trustworthiness (Trust.), attractiveness (Attract.), and dominance (Dom.) dissociate and also diverge from individual differences on identity recognition and expression labeling ability (shown on the off-diagonal, measured by Pearson's correlations). Data are from a nontwin participant group tested in the laboratory (n = 214, N observations = 1,070) (*SI Appendix*). (*C*) Best-fitting univariate AE twin models depicting variance explained in individual facial impressions from the twin sample, together with variance explained in individual facial attractiveness and facial identity recognition performance from different twin samples recruited from the same registry (12, 18). Hashing indicates upper-bound for estimates of additive genetic contribution, based on test–retest reliability.

variation in facial identity recognition ability, which is almost entirely genetically driven (Figs. 2*C* and 3*B*). The contrast between our present study of facial evaluation and previous studies on facial identity recognition is also especially striking given that the face-processing measures were reliable across these studies, and our sample was drawn from the same twin registry as used in a previous study of face identity recognition (12). Indeed, individual differences in facial impressions did not associate with either facial identity or emotional expression recognition ability (Fig. 2*B*).

Given that personal environments shaped all three facial impressions, a key question is whether the same or distinct environmental factors are responsible for driving individual variation across different impressions. The same environmental effect may be responsible for variation across the three impressions, based on theoretical accounts that suggest that individual differences in impressions may be shaped by people's general "diet" of faces, including by typicality (20), familiarity (33), and/or statistical learning (22). For example, faces that resemble those from one's own neighborhood, workplace, or school may receive generally more favorable impressions, contributing to individual differences across the three traits. Alternatively, distinct environmental factors may drive each impression. Models of facial impressions suggest that attractiveness, trustworthiness, and dominance can be dissociated at the group level (6, 7, 34). Here, we also found that individual differences in the three impressions were largely unrelated phenotypically (Fig. 2*B* and *SI Appendix*, Table S2). Distinct environmental factors would support more recent associative social learning theories which suggest that people learn to trust certain facial attributes based on specific previous social encounters (35, 36). For example, faces that resemble one's friends may be viewed as particularly trustworthy independent of the other impressions whereas faces that resemble one's romantic partners may be viewed as particularly attractive.

In order to address whether the same or distinct environmental factors were driving individual variation across different impressions, we modeled the trustworthiness, dominance, and attractiveness impressions simultaneously using multivariate modeling. Multivariate models divide the covariation between traits into A, C, and E factors, allowing estimates of the extent to which genetic and environmental influences on one impression are shared with the other impressions. In an iterative procedure, we compared full trivariate (Cholesky) models against more restrictive independent and common pathway models. The three types of model all include specific and shared variance for each impression but differ in how they partition the shared variance across impressions (*SI Appendix*, *Multivariate ACE modeling*). We also compared ACE models with simpler AE, CE, and E models of each type (*SI Appendix*, Tables S7–S9). The AE Cholesky model showed the best fit (as measured by AIC) (*SI Appendix*, Table S7) although model fit did not vary greatly between models. Multivariate heritability was complicated by negative genetic correlations, and not all confidence intervals converged. Nevertheless, the clearest effects across all models were substantial and significant effects for specific personal

Table 1. Reliability, Falconer's estimates, and variance component estimates for individual differences in trustworthiness, dominance, and attractiveness facial impressions (as measured by facial impression scores; see main text)

	Trustworthiness	Dominance	Attractiveness
Reliability			
Internal (split-half)	0.75	0.82	0.80
Test–retest	0.73	0.58	0.52
Falconer's formula: ACE estimates			
A	0.11	0.10	0.14
C	0.20	0.08	0.10
E	0.70	0.82	0.77
Model fit: −2 log-likelihood; AIC (*P* value for difference in −2 log-likelihood)			
ACE	3,016.32; 732.32	2,984.34; 718.34	2,781.98, 503.98
AE	3,016.71; 730.71, $P = 0.530$	2,984.34; 716.34, $P > 0.999$	2,781.98; 501.98, $P > 0.999$
CE	3,017.16; 731.16, $P = 0.359$	2,985.73; 717.73, $P = 0.238$	2,784.66; 504.66, $P = 0.102$
E	3,051.54; 763.54, $P < 0.001$	2,997.67; 727.67, $P = 0.001$	2,813.3; 531.3, $P < 0.001$
Full model: ACE estimates (95% CI)			
A	0.17 (0; 0.37)	0.18 (0; 0.27)	0.30 (0; 0.39)
C	0.11 (0; 0.33)	0 (0; 0.22)	0 (0; 0.30)
E	0.72 (0.63; 0.81)	0.82 (0.73; 0.92)	0.70 (0.61; 0.81)
Best-fit model: AE estimates (95% CI)			
A	0.29 (0.20; 0.38)	0.18 (0.08; 0.27)	0.30 (0.20; 0.39)
E	0.71 (0.62; 0.80)	0.82 (0.73; 0.92)	0.70 (0.61; 0.80)

A, additive genetic influences; C, shared environmental influences; E, both personal environmental influences and measurement error. Lower and upper bound 95% CIs are shown in parentheses for model estimates so that each estimate that does not touch zero is significant at alpha = 0.05 (two-sided). ACE, df = 1,142, 1,133, 1,139 for trustworthiness, dominance, attractiveness. difference in df ACE-AE = 1; difference in df ACE-CE = 1; difference in df ACE-E = 2.

environmental factors (i.e., not shared between impression types), mirroring the univariate impression models (e.g., 71 to 82% in the AE Cholesky model) (*SI Appendix*, Tables S8 and S9). This pattern is hard to explain with current familiarity, typicality, or statistical learning theories (20–22) because these theories are not context-specific: More familiar, typical, or normal faces should receive generally more favorable impressions, and, therefore, individual differences should be shared across the trait impressions. Instead, our results show that the environmental context matters for learning. This pattern supports a social learning model of individual differences in impressions, whereby observers learn to associate specific trait information with different facial cues (35, 36), but would also be consistent with an updated account of statistical learning which allows for contextual differences. The specificity of the environmental effects also rules out confounds, including a general halo or overall differences in scale use, which would occur across all impressions.

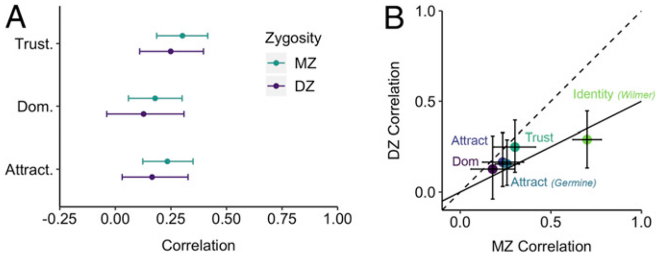

Fig. 3. Genetic and environmental contributions to face evaluation. (*A*) Correlations between MZ and DZ twin pairs on face impression scores for trustworthiness, dominance, and attractiveness, measured using intraclass correlation coefficients (ICC[1,1]) (32). Bars show 95% CIs; the large overlap between MZ and DZ correlations is indicative of a nongenetic contribution to individual differences in facial impressions. CIs were calculated using a bias-corrected and accelerated bootstrap with 2,000 bootstrap samples. (*B*) Plot of MZ vs. DZ intraclass correlations (ICCs) for the traits tested in the present study, as well as attractiveness face impressions (18) and Cambridge Face Memory Test face identity recognition ability (12, 13) reported elsewhere for twins recruited from the same registry. The dashed line represents the extreme case where all covariation within twin pairs is caused by shared environmental factors, resulting in equal ICCs for MZ and DZ pairs. The solid line represents the extreme case where all covariation within twin pairs is caused by additive genetic factors, such that MZ ICC is twice that for DZ pairs. Critically, in this case, the influence of both shared environmental factors and additive genetic factors was small relative to personal environmental factors. The further left and/or lower in the graph, the more idiosyncratic the judgments and the higher the contribution of personal environmental factors. All face evaluation results are well to the left of the face identity recognition result. Bars show 95% CIs.

Discussion

Here, we find large and stable individual variation in key facial evaluations of trustworthiness, dominance, and attractiveness, consistent with the classic idea that these visual judgments can be shaped by "the eye of the beholder." Using a twin study, we show that this variation in facial evaluation is largely shaped by people's personal experiences, rather than by genetic factors or shared environments. Highlighting the scope of personal experience to affect trust offers a different perspective on the fundamental basis, nature, and origin of individual trust and on our capacity to change whom we trust, for good or for ill. As our lives are increasingly affected by highly personalized social experiences, especially online (1, 2), our findings suggest that disagreements about whom we trust are also likely to increase.

Notably, our finding that variation in facial evaluation is driven by personal environments stands in sharp contrast to variation in facial recognition ability, which is almost entirely genetically driven (25). Multivariate modeling showed that the environmental factors driving individual differences in trustworthiness, dominance, and attractiveness evaluations were also largely independent. This pattern suggests that individual differences in impression formation are based on different experiences, and largely not based on overall or general familiarity, typicality, or overall statistical learning (20–22). Instead, our results are supportive of social learning theories, whereby unique social encounters shape individual associations between facial cues and associated traits (35, 36), or could also motivate new statistical learning theories which can account for the social context. Our results shed light on a core aspect of human social perception and indicate a remarkable diversity in the architecture of individual variation across different components of face processing.

As well as revealing the etiology of individual differences in trustworthiness and dominance evaluation, our results replicate and extend a behavioral genetics study of individual aesthetic judgments, which also found that individual differences in facial attractiveness are driven by people's personal experiences (18). Our current study used a new, more diverse (e.g., in age) and more naturalistic sample of faces. This demonstration of generalizability is especially critical here because the faces used will strongly affect the types of facial cues people can use to judge attractiveness and, consequently, available individual differences (9, 27).

Interestingly, our results do not necessarily imply that familial environment is unimportant even though the shared environment was not a major contributing factor. Siblings, including twins, can have remarkably unique familial environments (reviewed in ref. 29). For example, maternal affection can be very different even across identical twin pairs (29). Early caregiver or familial social experiences could therefore still influence unique mappings of facial cues to impressions.

Finally, it is important to be clear that our findings about individual differences do not argue against the claim that facial impressions of trustworthiness are adaptive, as suggested by leading facial impression theories (5–7, 26). Major evolutionary models of impressions have been based on consensus impressions (see ref. 17 for a review) whereas twin studies are concerned with individual variation. Facial cues that are critical for survival or successful reproduction may in fact be particularly strongly selected for, leading to consensus across individual perceivers. Indeed, consensus impressions, particularly of trustworthiness, are remarkably similar across cultural contexts, although there may be cultural "dialects" in impressions (37–39).

Our results suggest that a priority for future research should be to understand the development of social evaluation of faces. Especially, it will be critical to discover the developmental drivers of individual differences in face impressions, rather than focusing on potential genetic influences. We know little about how early in development these individual differences occur or which kinds of experiences are most consequential. One suggestion, based on our current findings, is that individual interactions with strangers, peers, and caregivers will be especially critical. A key methodological contribution of the current work is to provide a set of reliable tests of individual variation in trust and other impressions, which will benefit developmental and other research into individual differences in facial impression formation. As individual differences in facial impressions and identity recognition show distinctive etiologies, the perceptual and neural mechanisms driving variation in facial impressions will likely differ from those discovered in face recognition perception so far (reviewed in ref. 25). In terms of perceptual mechanisms, little is known about which facial features drive idiosyncratic impressions although a wealth of research has illustrated which facial features underlie consensus impressions (e.g., smiling, femininity, and raised eyebrow height are generally perceived as trustworthy) (6, 7). Idiosyncratic impressions could result from individually specific weighting of the same features that drive consensus trustworthiness impressions, as well as associations with additional features with trust or mistrust. Indeed, different facial features are likely to drive trustworthiness variation for different people, depending on their personal experiences (for example, one person may rely heavily on emotional expression to judge trustworthiness whereas another person relies on gender). Regarding neural mechanisms, plausible candidate neural regions driving individual impressions include the amygdala and caudate, which encode associative facial trust learning at the participant group level (23). Finally, the importance of individual experience, highlighted by our findings, motivates research to determine the long-term malleability of facial evaluations. This research aim is particularly critical, given the potential for these impressions to bias important social decisions, from online dating to courtroom sentencing (3, 17).

To conclude, we provide compelling evidence for substantial individual differences in impression formation and show that these differences are largely driven by unique personal environments, not genes (or shared environment). We also provide reliable tests of individual differences in impression formation. Our findings will speak to any scientist, philosopher, journalist, artist, or curious person who wonders why we judge a book by its cover, to what extent impressions lie in the eye of the beholder, and how our experiences with family, friends, partners, or the media might shape how we view the world.

Materials and Methods

Twin Participants. Twin participants were recruited from Twins Research Australia (40). Recruitment, testing, and analysis protocols were preregistered with Twins Research Australia and approved on 7 September 2017. Twin data collection was carried out from 1 June 2018 until 16 November 2018. Our final twin sample consisted of 1,264 individuals aged 16 to 80 y old (983 female; mean age, 47.4 y; SD age, 15.2 y), after a priori exclusions for inattention (see *SI Appendix*, *Twin sample* for more details). Participants were members of same-sex twin pairs (1,078 matched twin individuals and 186 unpaired twins).

For twin analyses, we excluded any participants with uncertain zygosity (*SI Appendix*, Table S3), leaving $n = 1{,}153$ consisting of 781 MZ individuals (568 female; mean age, 47.2 y; SD age, 15.2 y) and 372 DZ individuals (318 female; mean age, 49.8 y; SD age, 15.1 y). This subset included 490 matched twin pairs: 333 MZ pairs (241 female pairs; mean age, 48.1 y; SD age, 16.2 y) and 157 DZ pairs (135 female pairs; mean age, 49.9 y; SD age, 15.1 y).

The twin study was reviewed and approved by the Human Ethics Committee at the University of Western Australia and at Twins Research Australia. Participants gave informed consent before taking part in the study, which was conducted online. To incentivize careful participation, participants were given feedback about how their ratings compared to those of the average person (following ref. 41). Feedback was given after data collection was complete.

Measures.

Facial impression tests. Three tests measured individual differences on the three major dimensions of facial impressions: trustworthiness, dominance, and attractiveness. Psychometric test development of the three impression tests included extensive item selection (following ref. 18) (*SI Appendix*, *Test development and reliability*). The tests are openly available for future research. Test materials can be viewed online at https://www.testable.org/experiment/855/674205/start. Data and code are available in ref. 42.

In each test, participants rated a different set of 100 unknown male and female Caucasian faces, taken from a widely used, naturalistic face photograph database (43). Participants were initially informed of the trait to be rated and briefly viewed all 100 faces to familiarize them with the face variability. Participants rated the faces on the given trait from 1 ("Not at all trustworthy/dominant/attractive") to 9 ("Very trustworthy/dominant/attractive"). A subset of 50 faces was rated twice so that we could measure the consistency of participants' responses. Tests were self-paced, and each face remained on the screen until the response (median response time: 2.3 s). Participants completed the primary trustworthiness task first and then the control dominance and attractiveness tasks. Test development studies were reviewed and approved by the Human Ethics Committee at the University of Western Australia, and all participants provided informed consent. Data collection for test development studies was carried out between the 16 August 2017 and 8 May 2018.

Tests measure individual differences in facial impressions (i.e., the extent to which different people agree or not with consensus facial impressions of trustworthiness, dominance, or attractiveness) (Fig. 1). Specifically, scores on the tests represent a participant's overall agreement with average impressions on those faces, after controlling for unreliability in that participant's responses, so that differences in perception are not confounded by unreliability (following ref. 18). To quantify individual unreliability, we calculated the correlation between each participant's ratings of the subset of repeated faces across time points, resulting in a "self-consistency" score for each participant. The lower the "self-consistency" score, the higher the unreliability. We calculated each participant's agreement with the mean by correlating their individual ratings of the 100 faces (the first time they were seen) with the group mean ratings. We used a linear regression model to predict each participant's agreement score from their self-consistency score. The face impression scores are the residuals obtained from this model, which represent participants' agreement with the group mean, after controlling for individual participants' unreliability. High face impression scores reflect low idiosyncrasy.

Scenes control task. After the face tests, participants viewed a control task of 50 scenes (images taken from ref. 44), with a subset of 24 repeated scenes. Participants rated these scenes on their general preference ("how attractive do you find this scene?") from 1 (not at all attractive) to 7 (very attractive). This scene task acted as a control for individual variation in scale use, including participants' general idiosyncrasy, tendency to agree, and so on (18).

Zygosity and Demographics. Participants provided demographic information, including age and sex. Twin participants also gave us their birth order and completed an eight-item self-report questionnaire about their zygosity (45). Zygosity was determined using a latent class modeling approach (45).

Analysis. To compare twin models, we used −2 log-likelihood ratio tests (which approximate a distribution, with associated significance values to indicate a change in model fit) as well as Akaike's Information Criterion (AIC) (a lower AIC is better). We interpreted the size of genetic and environmental effects in line with previous effect size guidelines for twin modeling (29, 46). All CIs are 95% and all *P* values are two-tailed.

Data Availability. Data, code, and materials are available in ref. 42. The face impression tests can be viewed at https://www.testable.org/experiment/855/674205/start.

ACKNOWLEDGMENTS. This research was supported by Australian Research Council (ARC) Discovery Early Career Research Award DE190101043 (to C.A.M.S.), ARC Centre of Excellence in Cognition and Its Disorders Award CE110001021 (to G.R. and R.P.), and ARC Discovery Award DP170104602 (to G.R., C.A.M.S., and R.P.). We thank the staff at Twins Research Australia for their help with twin recruitment; staff at Testable for their help with data collection software; Scott Brindley, Jacob Hackett, and Joel Gray for help with pilot testing; and Neil Macrae for helpful comments on an earlier draft.

1. R. Botsman, *Who Can You Trust? How Technology Brought Us Together and Why it Might Drive Us Apart*, (Penguin, 2018).
2. W. Davies, *Nervous States: How Feeling Took Over the World*, (Vintage, 2019).
3. C. Y. Olivola, F. Funk, A. Todorov, Social attributions from faces bias human choices. *Trends Cogn. Sci.* **18**, 566–570 (2014).
4. J. P. Wilson, N. O. Rule, Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychol. Sci.* **26**, 1325–1331 (2015).
5. L. A. Zebrowitz, J. M. Montepare, Psychology. Appearance DOES matter. *Science* **308**, 1565–1566 (2005).
6. N. N. Oosterhof, A. Todorov, The functional basis of face evaluation. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 11087–11092 (2008).
7. C. A. M. Sutherland *et al*., Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition* **127**, 105–118 (2013).
8. R. J. W. Vernon, C. A. M. Sutherland, A. W. Young, T. Hartley, Modeling first impressions from highly variable facial images. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E3353–E3361 (2014).
9. E. Hehman, C. A. M. Sutherland, J. K. Flake, M. L. Slepian, The unique contributions of perceiver and target characteristics in person perception. *J. Pers. Soc. Psychol.* **113**, 513–529 (2017).
10. J. E. Martinez, F. Funk, A. Todorov, Quantifying idiosyncratic and shared contributions to judgment. *Behav. Res. Methods*, 10.3758/s13428-019-01323-0 (2020).
11. E. Hehman, R. M. Stolier, J. B. Freeman, J. K. Flake, S. Y. Xie, Toward a comprehensive model of face impressions: What we know, what we do not, and paths forward. *Soc. Personal. Psychol. Compass* **13**, e12431 (2019).
12. J. B. Wilmer *et al*., Human face recognition ability is specific and highly heritable. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 5238–5241 (2010).
13. N. G. Shakeshaft, R. Plomin, Genetic specificity of face recognition. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 12887–12892 (2015).
14. Q. Zhu *et al*., Heritability of the specific cognitive ability of face perception. *Curr. Biol.* **20**, 137–142 (2010).
15. T. J. Bouchard Jr., J. C. Loehlin, Genes, evolution, and personality. *Behav. Genet.* **31**, 243–273 (2001).
16. P. K. Hatemi, R. McDermott, The genetics of politics: Discovery, challenges, and progress. *Trends Genet.* **28**, 525–533 (2012).
17. A. Todorov, C. Y. Olivola, R. Dotsch, P. Mende-Siedlecki, Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annu. Rev. Psychol.* **66**, 519–545 (2015).
18. L. Germine *et al*., Individual aesthetic preferences for faces are shaped mostly by environments, not genes. *Curr. Biol.* **25**, 2684–2689 (2015).
19. C. A. M. Sutherland, A. W. Young, G. Rhodes, Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties. *Br. J. Psychol.* **108**, 397–415 (2017).
20. C. Sofer, R. Dotsch, D. H. J. Wigboldus, A. Todorov, What is typical is good: The influence of face typicality on perceived trustworthiness. *Psychol. Sci.* **26**, 39–47 (2015).
21. L. A. Zebrowitz, P. M. Bronstad, H. K. Lee, The contribution of face familiarity to ingroup favoritism and stereotyping. *Soc. Cogn.* **25**, 306–338 (2007).
22. R. Dotsch, R. R. Hassin, A. T. Todorov, Statistical learning shapes face evaluation. *Nat. Hum. Behav.* **1**, 1–6 (2016).
23. O. FeldmanHall *et al*., Stimulus generalization as a mechanism for learning to trust. *Proc. Natl. Acad. Sci. U S A* **115**, E1690–E1697 (2018).
24. V. Bruce, A. Young, Understanding face recognition. *Br. J. Psychol.* **77**, 305–327 (1986).
25. J. B. Wilmer, Individual differences in face recognition: A decade of discovery. *Curr. Dir. Psychol. Sci.* **26**, 225–230 (2017).
26. G. Rhodes, The evolutionary psychology of facial beauty. *Annu. Rev. Psychol.* **57**, 199–226 (2006).
27. J. Hönekopp, Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *J. Exp. Psychol. Hum. Percept. Perform.* **32**, 199–209 (2006).
28. M. C. Neale *et al*., OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika* **81**, 535–549 (2016).
29. R. Plomin, J. C. DeFries, V. S. Knopik, J. M. Neiderhiser, *Behavioral Genetics*, (Worth Publishers, ed. 5, 2008).
30. B. Duchaine, K. Nakayama, The cambridge face memory test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia* **44**, 576–585 (2006).
31. D. S. Falconer, T. F. C. MacKay, *Introduction to Quantitative Genetics*, (Longman's Green, ed. 4, 1996).
32. K. O. McGraw, S. P. Wong, Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* **1**, 30–46 (1996).
33. C. W. G. Clifford, G. Rhodes, *Fitting the Mind to the World: Adaptation and After-Effects in High-Level Vision*, (Oxford University Press, 2005).
34. M. Walker, T. Vetter, Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *J. Vis.* **9**, 1–13 (2009).
35. S. C. Verosky, A. Todorov, When physical similarity matters: Mechanisms underlying affective learning generalization to the evaluation of novel faces. *J. Exp. Soc. Psychol.* **49**, 661–669 (2013).
36. R. Hassin, Y. Trope, Facing faces: Studies on the cognitive aspects of physiognomy. *J. Pers. Soc. Psychol.* **78**, 837–852 (2000).
37. L. A. Zebrowitz *et al*., First impressions from faces among US and culturally isolated Tsimane' people in the Bolivian rainforest. *J. Cross Cult. Psychol.* **43**, 119–134 (2012).
38. C. A. M. Sutherland *et al*., Facial first impressions across culture: Data-driven modeling of Chinese and British perceivers' unconstrained facial impressions. *Pers. Soc. Psychol. Bull.* **44**, 521–537 (2018).
39. C. A. M. Sutherland, G. Rhodes, N. S. Burton, A. W. Young, Do facial first impressions reflect a shared social reality? *Br. J. Psychol.* (2019).
40. J. L. Hopper, The Australian twin registry. *Twin Res.* **5**, 329–336 (2002).
41. L. Germine *et al*., Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychon. Bull. Rev.* **19**, 847–857 (2012).
42. C. A. M. Sutherland *et al*, Individual first impressions from faces are shaped mostly by environments, not genes. Open Science Framework. https://osf.io/35zf8/?view_only=e76c6755dcea4be2adc5b075cae896e8. Deposited 18 April 2019.
43. W. A. Bainbridge, P. Isola, I. Blank, A. Oliva, "Establishing a database for studying human face photograph memory" in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, N. Miyake, D. Peebles, R. P. Cooper, Eds. (Cognitive Science Society, Austin, TX, 2012), pp. 1302–1307.
44. E. A. Vessel, N. Rubin, Beauty and the beholder: Highly individual taste for abstract, but not real-world images. *J. Vis.* **10**, 1–14 (2010).
45. A. C. Heath *et al*., Zygosity diagnosis in the absence of genotypic data: An approach using latent class analysis. *Twin Res.* **6**, 22–26 (2003).
46. S. Clifford, K. Lemery-Chalfant, H. H. Goldsmith, The unique and shared genetic and environmental contributions to fear, anger, and sadness in childhood. *Child Dev.* **86**, 1538–1556 (2015).