The structure and perceptual basis of social judgments from faces

Alexander Todorov[a,*] **and DongWon Oh***[,b]
[a]The University of Chicago Booth School of Business, Chicago, IL, United States
[b]Department of Psychology, New York University, New York, New York, United States
*Corresponding author: e-mail address: alexander.todorov@chicagobooth.edu; dongwon.oh@nyu.edu

Contents

Abstract

Two questions have motivated much of our research. What is the structure of social judgments from faces and what is the perceptual basis of these judgments? Twelve years ago, we proposed a simple 2-dimensional model, according to which faces are evaluated on perceived valence and power, and introduced data-driven computational models, which reveal the physical variation of faces that drive specific judgments. First, using data from our lab collected more than 10 years later and from a large cross-cultural replication project, we show that the simplistic 2D structure model is remarkably stable and generalizes to all world regions, in which data have been collected. We also discuss misunderstandings and limitations of structure models. Second, we make the assumptions of the data-driven modeling approach transparent and discuss recent developments. The computational framework allows for precise parametric manipulation of the appearance of faces, for control of shared variance between judgments, and for relating models at different levels of face processing. We also explore analyses mapping

Advances in Experimental Social Psychology
ISSN 0065-2601
https://doi.org/10.1016/bs.aesp.2020.11.004

multiple social judgments simultaneously to the physical face space. We conclude with two future directions: models of hyper-realistic faces and the underappreciated role of stable individual differences in judgments.

Face perception has long been a key area of research in psychology and neuroscience, because this research often addresses fundamental questions about the evolution, development, and organization of the mind. Those questions include whether we are born with innately specified biases to process stimuli such as faces, the role of experience in tuning our exquisite perceptual abilities, and whether the mind is organized as a set of inter-locked domain specific modules or has a domain general architecture (see Chapters 12 & 13 of Todorov, 2017).

However, in contrast to many fields in psychology—cognitive, developmental, and evolutionary—and cognitive neuroscience (Calder, Rhodes, Johnson, & Haxby, 2011) where face perception research has been prominent for decades, research on faces was not part of mainstream social psychology until recently. One exception is the long research tradition of studying emotional expressions (Barrett, Adolphs, Marsella, Martinez, & Pollak, 2019), though much of this research would be classified today as belonging to affective science. Of course, the fact that face research was not part of mainstream social psychology does not mean that there were no systematic research programs in the field. Two prominent examples are the early work of Paul Secord in the 1950s (e.g., Secord, 1958), and the subsequent seminal work of Leslie Zebrowitz starting in the 1980s and continuing until today (e.g., Berry & Zebrowitz McArthur, 1985; Montepare & Zebrowitz, 1998; Zebrowitz, 1997, 2017; Zebrowitz & Collins, 1997).

The absence of faces from mainstream social psychology was surprising given that faces are rich ecological stimuli triggering the sorts of inferences that have always been of interest to social psychologists: from belonging to particular social groups and the associated stereotypes with these groups (Dotsch, Wigboldus, Langner, & van Knippenberg, 2008; Dotsch, Wigboldus, & van Knippenberg, 2011; Freeman, Rule, Adams, & Ambady, 2010; Johnson, Freeman, & Pauker, 2012; Ma, Correll, et al., 2018; Ma, Koltai, et al., 2018; Macrae & Bodenhausen, 2000) to inferences about emotional and mental states, and even personality (Todorov, 2017). Moreover, these inferences are rapidly made from minimal information (Ballew & Todorov, 2007; Bar, Neta, & Linz, 2006; Colombatto, Uddenberg, & Scholl, under review; Todorov, Loehr, & Oosterhof, 2010; Todorov, Pakrashi, & Oosterhof, 2009; Willis & Todorov, 2006) and are present early in

development (Charlesworth, Hudson, Cogsdill, Spelke, & Banaji, 2019; Cogsdill & Banaji, 2015; Cogsdill, Todorov, Spelke, & Banaji, 2014; Jessen & Grossmann, 2016). Further, these inferences are not only influenced by bottom up facial cues, but are also shaped by perceivers' socio-conceptual knowledge (e.g., preconceptions about social groups, emotions, and traits) (Freeman, Stolier, & Brooks, 2020).

Our informal assessment of the field today is that all this has changed and research on face perception is thriving in social psychology. Moreover, the disciplinary boundaries of face perception research have been blurred and researchers from various disciplines address similar questions that are not artificially relegated to one or another field (Calder et al., 2011; Young, 2018). The research we have conducted over the last 15–20 years is one small contribution to this inter-disciplinary trend.

The interest of the first author in faces started during his graduate school days (Todorov & Uleman, 2002, 2003, 2004). An old research debate in the literature on spontaneous trait inferences (e.g., honesty) was about whether such inferences from behavioral statements (e.g., returned the lost wallet) get associated with the person enacting the behavior. It turned out that such associations are easily detected if the person is instantiated with an actual face rather than a personal name or a label describing an occupation (Uleman, Blader, & Todorov, 2005). We have continued this research over the years (Falvello, Vinson, Ferrari, & Todorov, 2015; FeldmanHall et al., 2018; Ferrari, Oh, Labbree, & Todorov, 2020; Goren & Todorov, 2009; Todorov, Gobbini, Evans, & Haxby, 2007; Verosky, Porter, Martinez, & Todorov, 2018; Verosky & Todorov, 2010, 2013; Verosky, Todorov, & Turk-Browne, 2013), though it has received much less attention than our work on social attributions from facial appearance (Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). In fact, people are remarkably good at forming affective associations with faces based on relevant behavioral information (Falvello et al., 2015; Ferrari et al., 2020). These findings are good to remember in light of the research presented below on shallow but rapid inferences from facial appearance. These inferences are easily modified when people are provided with information about past actions.

Todorov's interest in appearance-based inferences really peaked after two sets of findings. The first was that naïve judgments of competence based solely on the facial appearance of politicians predicted the outcomes of important elections in the US (Todorov, Mandisodza, Goren, & Hall, 2005). Since then there have been many replications, including several conducted in countries other than the US (Antonakis & Dalgas, 2009;

Lawson, Lenz, Baker, & Myers, 2010; Olivola & Todorov, 2010a; Poutvaara, Jordahl, & Berggren, 2009; Sussman, Petkova, & Todorov, 2013), and political scientists have figured out the potential mechanisms underlying the effects of appearance (Ahler, Citrin, Dougal, & Lenz, 2017; Lenz & Lawson, 2011): appearance primarily influences uninformed voters who have been exposed to images of political candidates. This research demonstrated the importance of shallow first impressions for real world outcomes (see Chapter 3 of Todorov, 2017).

The second set of findings was that people can make social judgments after extremely brief exposures to faces (Willis & Todorov, 2006). In our first studies, participants saw faces for 100, 500, or 1000 milliseconds. Contrary to our expectations, the nature of the judgments did not change with increased exposure. Longer exposures only increased confidence in judgments. One problem with these initial studies was that the faces were not properly masked and, hence, the actual exposure to faces may have been longer than the intended exposures. Many subsequent studies have ruled out this explanation of the findings (Bar et al., 2006; Todorov et al., 2009; Todorov et al., 2010). Now we know that as little as 50 milliseconds exposure to faces provides sufficient information for people to make reliable social judgments (i.e., *consistent* judgments over time, not to be confused with *accurate* judgments of others based on their facial appearance), and that these judgments do not change with exposures longer than about 200 ms.

These two sets of findings—on the importance of shallow judgments from faces and the efficiency of these judgments—led to more systematic research on social judgments from faces. It also helped that the first author was urged by senior colleagues at the time to develop a systematic research program. The first key paper of our lab reporting on this program was *The Functional Basis of Face Evaluation* (Oosterhof & Todorov, 2008). Here we are revisiting this work 12 years later.

The Functional Basis of Face Evaluation was motivated by two research questions. The first question was: is there a simple structure underlying social judgments from faces, given their high inter-correlations (Fig. 1A; see also Fig. 2)? The second question was: what is the perceptual basis of these judgments, given that there is agreement among people when they evaluate faces (Fig. 1B)? These two questions have continued to motivate much of our research. It should be noted that both questions deal with the construction of social judgments, *not* with their accuracy (for the low accuracy or validity of social judgments from faces, see part 3 of Todorov, 2017; as well as Olivola & Todorov, 2010b; Todorov et al., 2015).

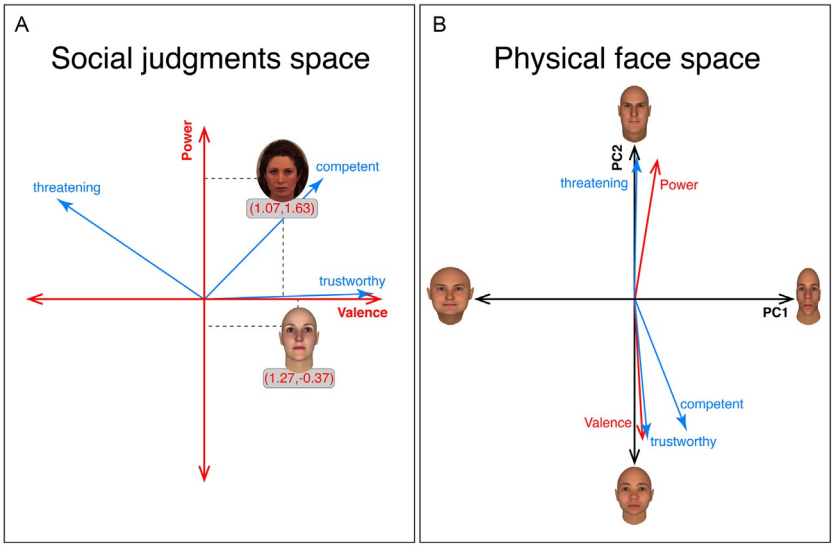

Fig. 1 Face spaces of social judgments and physical properties. (A) The social judgment space is extracted from judgments of faces, and illustrates relations between judgments. Geometrically, the correlation between two vectors (judgments here) is given by the cosine of the angle between them. The smaller the angle, the higher the correlation is. In the social judgment space, it could be seen that judgments of competence are positively correlated with judgments of trustworthiness and are almost orthogonal to judgments of threat; the latter are negatively correlated with judgments of trustworthiness. Any face can be represented in this space and two examples are shown (the coordinates are the faces' principal component scores from Oosterhof & Todorov, 2008). (B) In contrast to the social judgment space, the physical face space is extracted from a statistical analysis of face variation. Typically, physical spaces have many more dimensions than judgment spaces. Because of this, the angles between the vectors here are only informative about their correlation in the first 2 dimensions of the model. For the actual correlations between models of judgments in a 100-dimensional space, see Fig. 13.

Although these two questions are related, they are both conceptually and empirically independent. Finding a social judgment space does not require a model of the physical representation of faces. One only needs multiple social judgments of faces. Finding a simple structure underlying judgments does imply a potential mapping between this structure and the physical representation of faces, but the latter is not necessary in the search for the former. Similarly, one can try to model any social judgment in the physical face space irrespective of how this judgment relates to other social judgments. The judgment and physical face spaces are two distinct spaces extracted from human judgments and a statistical analysis of face variation, respectively (O'Toole, 2011). However, a comprehensive research program should try to find a mapping between these two spaces (Over & Cook, 2018).

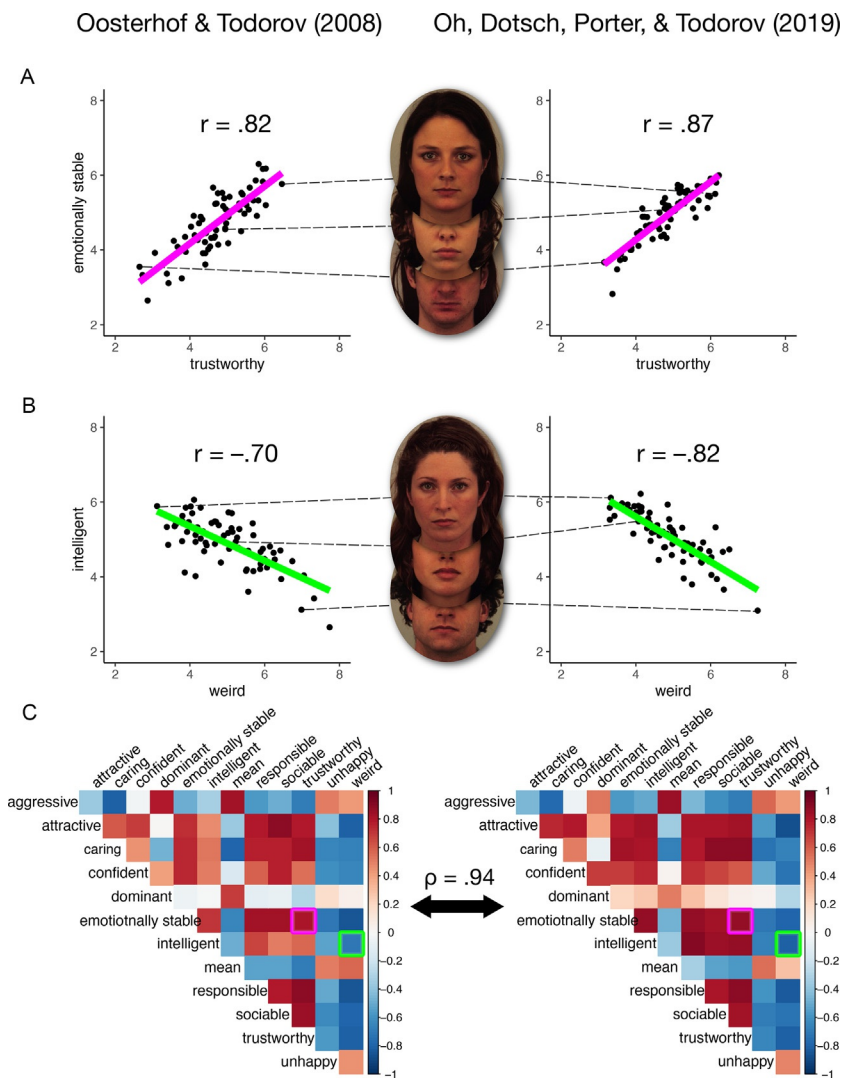

Fig. 2 Pairwise correlations of trait judgments from 2008 and 2019 datasets. (A) The correlation between judgments of trustworthiness and emotional stability. Each point in the scatter plots is a face. (B) The correlation between judgments of weirdness and intelligence. (C) Pearson correlational coefficients for all pairs of 13 social judgments ($n = 78$) in each dataset, represented as two matrices. The degree of similarity between these two matrices is assessed by computing a non-parametric Spearman correlation, which only assumes a monotonic relationship between the vectors of Pearson coefficients. The correlational structures in the two datasets are highly similar to each other.

In *The Functional Basis of Face Evaluation*, we tried to accomplish two things. The first was to offer a systematic way of organizing social judgments from faces in a simplistic 2D model. We refer to these kinds of models as *dimensional structure models*—models that are derived from an analysis of the similarity of social judgments (Fig. 1A). The first part of this chapter is about these models. Using data from our lab collected more than 10 years later (Oh, Dotsch, Porter, et al., 2019) and from a large cross-cultural replication project (Jones et al., 2020), we show that the simplistic 2D model that we offered in 2008 is remarkably stable and generalizes to all 11 world regions (41 countries, $N > 11{,}000$), in which data were collected. We also try to dispel some misconceptions about the interpretation of dimensional structure models.

The second thing we tried to accomplish was to introduce data-driven computational methods that can discover the perceptual basis of social judgments and offer means for systematic manipulation of the appearance of faces. We refer to the resulting models of judgments as *computational models of judgments*—models that are extracted from modeling a specific judgment (e.g., trustworthiness) or a set of judgments as a function of variations in the physical face space (Fig. 1B) and that can parametrically manipulate the appearance of faces. The second part of this chapter is about these computational models. We try to make the assumptions of the methods as clear as possible, describe multiple models of social judgments, show how the methods extend to measures other than explicit judgments, and explore analyses mapping multiple social judgments simultaneously to the physical face space.

We conclude the chapter with two future directions of research. First, we speculate that the computational models we introduced are going to be replaced by models based on deep neural networks (e.g., Karras et al., 2019), but in the process we will trade off hyper-realism of faces for conceptual clarity and transparency. Second, we discuss the largely overlooked topic of stable individual differences in social judgments of faces (Hönekopp, 2006; Martinez, Funk, & Todorov, 2020), differences that are masked by the fact that many models are based on aggregated judgments.

1. The robust beauty of the structure of judgments from faces

In Section 1.1, using both a principal component analysis (PCA) and an exploratory factor analysis (EFA), we identify a simple two-dimensional

structure underlying social judgments of faces. We also show that this structure is highly stable using data collected more than 10 years apart. Section 1.2 is about the universality of the structure of social judgments. Capitalizing on a recent large cross-cultural replication project (Jones et al., 2020), we show that this structure is found across all world regions, in which data were collected. In Section 1.3, we try to clear out misconceptions about dimensional structure models and discuss their implications for cross-cultural differences. Finally, in Section 1.4, we discuss the limitations of dimensional structure models.

1.1 The structure of social judgments from faces

As mentioned earlier, social judgments from faces are highly correlated with each other. Fig. 2A and B illustrate these high correlations for two pairs of judgments: trustworthy and emotionally stable, and weird and intelligent. The data for these judgments were collected more than 10 years apart (Oh, Dotsch, Porter, et al., 2019; Oosterhof & Todorov, 2008). In both years of data collection, the correlations are high and similar to each other. Faces perceived as trustworthy are also perceived as emotionally stable. Faces perceived as weird are also perceived as less intelligent. The magnitude of the correlations is above .70. Fig. 2C shows the correlational structure for all pairs of judgments. Two things should be noted: most pairs of judgments are highly correlated and the correlational structures for the two datasets are highly similar ($\rho = 0.94$).

The high correlations between social judgments of faces motivated the first question of our research program: finding the structure of these judgments (Oosterhof & Todorov, 2008). To find this structure, one must rely on dimensionality reduction techniques, such as PCA. However, any data reduction technique is completely dependent on the data input. In this specific case, given that there are thousands of trait adjectives and potential judgments, what should be the traits on which faces are judged? Rather than relying on our own hunches, we collected unconstrained descriptions of faces and reduced these descriptions to traits. Most of the descriptions were related to 12 traits (see Table 1). We also included dominance because of the importance of this trait in prior models of social judgments (Wiggins, 1979; Wiggins, Phillips, & Trapnell, 1989). After we identified the relevant traits, we asked participants to rate faces on each of the 13 selected traits (13 different groups of participants). The average absolute correlation between the judgments was .58. Given these high inter-correlations, we submitted the

Table 1 Principal component loadings of social judgments of the same set of faces from two different waves of data collection.

	2008[a]		2019		2019[b]	
Judgment	**PC 1**	**PC 2**	**PC 1**	**PC 2**	**PC 1**	**PC 2**
Aggressive	**−0.69**	**0.68**	**−0.65**	**0.70**	**−0.81**	**0.65**
Attractive	**0.81**	0.30	**0.89**	0.24	**0.82**	**0.31**
Caring	**0.89**	−0.29	**0.92**	−0.26	**0.97**	−0.19
Confident	**0.67**	**0.64**	**0.73**	**0.62**	**0.57**	**0.68**
Dominant	−0.26	**0.93**	0.17	**0.93**	−0.06	**0.96**
Emotionally stable	**0.93**	0.18	**0.93**	0.04	**0.91**	0.12
Intelligent	**0.71**	0.16	**0.93**	0.15	**0.88**	0.23
Mean	**−0.74**	**0.56**	**−0.55**	**0.74**	**−0.71**	**0.70**
Responsible	**0.90**	0.12	**0.92**	0.19	**0.86**	0.26
Sociable	**0.90**	0.21	**0.92**	−0.03	**0.92**	0.04
Trustworthy	**0.93**	−0.06	**0.95**	−0.10	**0.95**	−0.02
Unhappy	**−0.71**	0.00	**−0.76**	0.19	**−0.79**	0.13
Weird	**−0.88**	−0.21	**−0.86**	−0.23	**−0.79**	−0.30
Explained variance	63%	18%	66%	20%		

[a]An observant reader might notice that the values reported here are slightly different from the values reported in Oosterhof and Todorov (2008, table S3). The reason is that the original PCA was conducted on the averages of the standardized scores of individual participants' judgments (see p. 1 of SM in Oosterhof & Todorov, 2008), not the average of the unstandardized scores. For consistency with subsequent studies (including the cross-cultural replication described below), we use the latter here.
[b]The loadings in these two columns are computed after a Procrustes rotation to align the 2019 solution to the 2008 solution. See text for details.
The data were collected more than 10 years apart (Oh, Dotsch, Porter, et al., 2019; Oosterhof & Todorov, 2008). Loadings greater than 0.30 are in bold font.

ratings to a PCA. The PCA identified two components, which we interpreted as valence/trustworthiness and power/dominance (Oosterhof & Todorov, 2008; Todorov, Said, Engell, & Oosterhof, 2008).

To test the stability of the two-dimensional structure of social judgments, we directly compare the original PCA findings (as well as EFA findings) to findings from data collected more than 10 years later. As a part of a recent project (Oh, Dotsch, Porter, et al., 2019), we collected the same trait ratings of the faces we used in 2008. The average absolute correlation between the judgments was high (0.61) and comparable in size to the average absolute

correlation in 2008. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy, indicating the suitability of data for dimensionality reduction, was very high (0.88 for 2008 and 0.91 for 2019). As shown in Table 1, the PCA solutions were highly similar. This should not be surprising, given the high similarity of the correlational structures (see Fig. 2C).

In both datasets, the first principal component (PC) accounts for more than 60% of the variance. All positive judgments (e.g., caring) have positive loadings and all negative judgments (e.g., aggressive) have negative loadings, suggesting that this component can be interpreted as valence. The only judgment that is not highly correlated with this component is dominance. In both datasets, the second PC accounts for more than 18% of the variance. Judgments of dominance have the highest loadings followed by judgments of aggressiveness and confidence, suggesting that this component can be interpreted as power. One difference between the solutions is that whereas dominance is negatively loaded on the first PC in 2008, it is positively loaded in 2019. It is likely that this difference has to do with differences in the samples of respondents: Princeton students in 2008 vs. M-Turk participants in 2019. In any case, the loadings are small in magnitude.

We can also compare the two PCA solutions quantitatively by computing factor congruence coefficients (Abdi, 2007; Tucker, 1951). These coefficients indicate the similarity of principal components or factors. Typically, coefficients above 0.95 indicate that the two PCs or factors are the same. Coefficients between .85 and .95 indicate that they are fairly similar (Lorenzo-Seva & Ten Berge, 2006). Before computing factor congruence coefficients, it is important to align the two loading matrices by a Procrustes rotation, which rotates the axes of one of the matrices to align with the other while preserving the pairwise distances between points in each dataset (Fischer & Fontaine, 2011; Fischer & Karl, 2019). This procedure is particularly important in the case of EFA, because there are infinitely many factor solutions that are mathematically equivalent (a topic that we revisit later).

Table 1 shows the rotated loadings of the judgments from the 2019 dataset after alignment to the 2008 dataset. The congruence coefficient is 0.993 for the first PC and 0.978 for the second PC. Even without a Procrustes rotation, the coefficients are very high (0.983 and 0.968 for the two PCs, respectively). Thus, we can conclude that the structure of the judgments is the same across these two datasets collected more than 10 years apart.

It is important to explain why we chose two as opposed to three principal components. There are multiple soft (but no exact) rules on the selection of

the number of components. For both datasets, the third PC accounted for less variance than a single variable (the formal justification of the Kaiser rule to retain only components with eigenvalues greater than 1). More importantly, in contrast to the first two PCs, the third PC was not easily interpretable. The only judgment with high loading on this PC was unhappy (0.61); the judgment with the next highest absolute loading was weird, but with opposite sign (−0.28). Ultimately, interpretability, though subjective, is the final criterion for retaining components or factors. No component or factor comes with a ready label.

It is also important to explain why we chose to submit the data to PCA as opposed to EFA in our initial analyses. PCA is one of the simplest techniques for dimensionality reduction, making minimal assumptions. The first PC is the linear combination of the input variables that accounts for more of their variance than any other linear combination. The second PC is the next linear combination, which is orthogonal to the first that accounts for more variance than other linear combinations, and so on. The objective of the analysis is data reduction: to account for as much variance with as few components as possible.

EFA is typically heralded as a superior technique revealing the latent structure of the data. However, EFA solutions are always underdetermined by the data. There are two types of well-known indeterminacies: due to the estimation of the commonality problem and due to the factor rotation problem (Sharma, 1996). Because the objective of EFA is to identify latent factors that explain the correlations among the input variables, for each variable this analysis has to estimate both the common variance with the latent factors and the unique variance of the variable. This estimation is always underdetermined: to estimate the common variance, one needs to know the unique variance, but to know the unique variance, one needs to know the common variance. The second indeterminacy arises from the rotation problem. Unlike PCA, the axes in EFA are rotated. The projection of the variables on these rotated axes determines their commonality, but there is an infinite number of rotations that would decompose the same commonality differently. Because of these indeterminacies, there is no unique factor solution, but infinitely many solutions compatible with the same dataset.

In our initial analyses (Oosterhof & Todorov, 2008; Todorov et al., 2008), we erred on the side of simplicity and chose to submit the data to PCA rather than EFA. We were also interested in dimensionality reduction, given the high correlations among judgments, rather than in revealing latent factors. Nevertheless, here we also analyze the 2008 and 2019 data using EFA.

Table 2 Factor pattern loadings of social judgments of the same set of faces from two different waves of data collection.

	2008		2019		2019[a]	
Judgment	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**
Aggressive	−0.30	**0.85**	**−0.50**	**0.77**	**−0.45**	**0.80**
Attractive	**0.86**	0.08	**0.92**	0.13	**0.92**	0.07
Caring	**0.67**	**−0.50**	**0.85**	**−0.36**	**0.82**	**−0.41**
Confident	**0.91**	**0.46**	**0.84**	**0.54**	**0.88**	**0.48**
Dominant	0.22	**0.98**	**0.34**	**0.89**	**0.41**	**0.86**
Emotionally stable	**0.93**	−0.05	**0.92**	−0.05	**0.92**	−0.11
Intelligent	**0.68**	−0.03	**0.94**	0.05	**0.95**	−0.01
Mean	**−0.40**	**0.71**	**−0.39**	**0.75**	**−0.33**	**0.77**
Responsible	**0.87**	−0.10	**0.93**	0.09	**0.94**	0.03
Sociable	**0.92**	−0.01	**0.89**	−0.13	**0.88**	−0.19
Trustworthy	**0.82**	−0.29	**0.91**	−0.20	**0.89**	−0.25
Unhappy	**−0.62**	0.16	**−0.68**	0.24	**−0.66**	0.28
Weird	**−0.89**	0.01	**−0.87**	−0.12	**−0.88**	−0.07
Explained variance	56%	23%	64%	20%		

[a]The loadings in these two columns are computed after a Procrustes rotation to align the 2019 solution to the 2008 solution. See text for details.
The data were collected more than 10 years apart (Oh, Dotsch, Porter, et al., 2019; Oosterhof & Todorov, 2008). Loadings greater than 0.30 are in bold font.

The estimation procedure was ordinary least squares, which finds the minimum residuals solution. We used an oblimin rotation, which allows for the factors to be correlated. As shown in Table 2, the factor solutions were very similar.

In both datasets, all positive judgments (e.g., caring) have positive loadings and all negative judgments (e.g., mean) have negative loadings on the first factor, suggesting that it can be interpreted as valence. In both datasets, judgments of dominance have the highest loadings on the second factor followed by judgments of aggressiveness and meanness, suggesting that this factor can be interpreted as power or possibly threat. As in the case of the PCA, quantitatively, the factors are also structurally identical. The congruence coefficients for the two factors are 0.991 and 0.982, respectively. Again, the coefficients are very high even without a Procrustes rotation (0.991 and 0.976, respectively).

Comparing the EFA solutions (Table 2) to the PCA solutions (Table 1) suggests that the second factor (relative to the second PC) acquires a more evaluative, negative meaning. For example, confidence judgments are much more highly loaded on the first than on the second factor. Aggressiveness and meanness judgments have higher loadings than confidence on the second factor, and unambiguously positive judgments such as caring and trustworthy increase their negative weight on this factor (relative to their loadings on the second PC). This interpretation is consistent with the negative correlation between the two factors: −0.25 for 2008 and −0.10 for 2019. Finally, in 2019 we also collected judgments of threat and can correlate these judgments with the factor and the PC scores of the faces. For both datasets, these judgments were more highly correlated with the second factor (0.72 for 2008, 0.63 for 2019) than with the second PC (0.42 for 2008, 0.48 for 2019).

To summarize, the structure of social judgments is stable and can be described by two dimensions: valence and power. Although the second dimension acquires a more negative meaning when the dimensions are extracted by an EFA, the PCA and EFA solutions are similar.

1.2 The universality of the structure of social judgments from faces

Recently, the Psychology Science Accelerator project (PSA; Moshontz et al., 2018) initiated a large cross-cultural replication of our two-dimensional model (Jones et al., 2020). Respondents ($N = 11{,}481$) from 11 world regions (see Table 4) and 41 countries, covering the whole world, rated faces on the same 13 traits we used in our original research. In this section, we test how well the 2D model holds in all world regions. Given the structural invariance of the 2008 and 2019 data, we average across the two datasets (see Table 3) and compare this aggregated dataset with the data from each of the 11 world regions.

On a priori grounds, it is easier to predict that the 2D model would not hold across all cultures (Henrich, Heine, & Norenzayan, 2010). Much research in psychology has been criticized on the grounds that it is Western-centered and there have been studies suggesting cultural differences in face evaluation (Han et al., 2018; Nakamura & Watanabe, 2019; Scott et al., 2014; South Palomares, Sutherland, & Young, 2018; Sutherland et al., 2018; Zhang et al., 2019). Thus, theoretically one would expect substantial geographical and cultural variation. Methodologically, the different languages and subtle differences in the meaning of trait attributes in different cultures would all introduce additional variation. Moreover, the set of faces used by the PSA researchers was larger and more diverse than our set. We used the Karolinska

Table 3 Principal component and factor pattern loadings of social judgments of faces derived from a principal component analysis (PCA) and an exploratory factor analysis (EFA).

	PCA		EFA	
Judgment	**PC 1**	**PC 2**	**Factor 1**	**Factor 2**
Aggressive	**−0.69**	**0.69**	**−0.42**	**0.80**
Attractive	**0.86**	0.29	**0.90**	0.12
Caring	**0.91**	−0.27	**0.77**	**−0.42**
Confident	**0.70**	**0.65**	**0.88**	**0.52**
Dominant	−0.09	**0.97**	0.25	**0.98**
Emotionally stable	**0.95**	0.11	**0.94**	−0.05
Intelligent	**0.86**	0.16	**0.85**	0.00
Mean	**−0.67**	**0.67**	**−0.42**	**0.75**
Responsible	**0.93**	0.17	**0.94**	0.01
Sociable	**0.95**	0.11	**0.94**	−0.06
Trustworthy	**0.96**	−0.07	**0.89**	−0.23
Unhappy	**−0.77**	0.08	**−0.68**	0.19
Weird	**−0.88**	−0.21	**−0.90**	−0.05
Explained variance	67%	20%	63%	22%

The data for the analyses were judgments aggregated across two datasets (Oosterhof & Todorov, 2008; Oh, Dotsch, Porter, et al., 2019). Loadings greater than 0.30 are in bold font.
The correlation between the two factors for the EFA is −0.18.

database (Lundqvist, Flykt, & Öhman, 1998), which is comprised of photographs of white individuals. The PSA project used the much more diverse Chicago database (Ma, Correll, & Wittenbrink, 2015) and selected faces representing four different ethnicities (Asian, Black, Latino, and White). The differences between the stimuli used should also introduce additional variation.

However, as shown in Fig. 3, the correlational structure of judgments around the world is remarkably similar to the correlational structure of our dataset. For every region, the pattern of pairwise correlations between judgments is similar to the pattern of our data collected on a different and much less diverse set of (all White) faces. The lowest correlation is with the data from the Middle East (0.88). For any other region, the correlation is equal to or above 0.92.

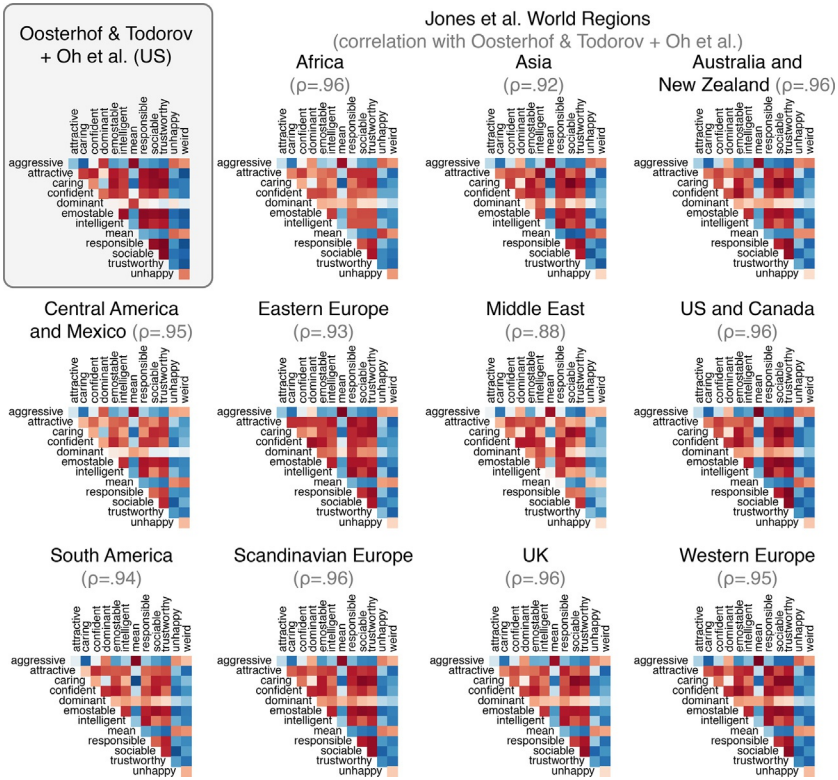

Fig. 3 The stability of the correlational structure of social judgments from faces across the world. Pearson correlational coefficients for all pairs of 13 social judgments ($n = 78$) in datasets from Todorov lab (derived from the mean ratings averaged between Oosterhof & Todorov, 2008 and Oh, Dotsch, Porter, et al., 2019) and datasets from 11 world regions (Jones et al., 2020). Despite the fact that the judgments for the Todorov lab dataset were made on a completely different set of faces, their correlational structure is remarkably similar to the correlational structure of the judgments in each of the 11 world regions. The similarity is assessed by computing non-parametric Spearman correlations.

Following the steps of Jones et al. (2020), to quantitatively compare the PCA and EFA solutions for our own data with the corresponding solutions for each world region, we conducted a PCA and an EFA for each world region and then computed the factor congruence coefficients between the respective PCs (for the PCA) and the respective factors (for the EFA). However, there are three important differences. First, a priori, we decided on extracting two-dimensional solutions, because these are easily interpretable (see the previous section), but we also discuss three-dimensional solutions at the end of this section. Moreover, when comparing EFA solutions

from different datasets, the data should be in the same dimensional space. Unlike PCA, where extracting additional PCs does not change the variables' loadings on the already extracted PCs (e.g., the loading matrix for the first two PCs remains the same whether one extracts 2, 5, or 13 PCs), in EFA extracting additional factors does change the loadings on the already extracted factors. Thus, comparing a two-factor solution with a three-factor solution is problematic. Second, before computing the congruence coefficients, we aligned each of the world region matrices to the matrix of our dataset using a Procrustes rotation (see the previous section). We also ran bootstrapping simulations to show that the observed high congruence between our data and world region data is extremely unlikely to be an artifact of the Procrustes procedure. Third, we used as a benchmark of comparison the combined dataset from Oosterhof and Todorov (2008) and Oh, Dotsch, Porter, et al. (2019). This last difference is the least important and the results do not change if each of the dataset is used as the benchmark of comparison. However, given the structural invariance of the two datasets and for simplicity (see the previous section), we used the combined dataset (Table 3).

Table 4 shows that the structure of judgments derived from both a PCA and an EFA is remarkably consistent across cultures. This should not be surprising in light of Fig. 3. After all, the input to these analyses is the correlations between judgments. As shown in Table 4, for the first PC derived from the PCA, the range of the factor congruence coefficients is from 0.975 to 0.995. That is, this component would be considered structurally identical in all world regions. For the second PC, the range is from 0.885 to 0.989. This component would be considered identical in nine world regions and very similar in the remaining two regions (Asia and the Middle East).

The Procrustes rotation has been criticized on the grounds that it might inflate the similarity between factor matrices, even when these are randomly generated (Paunonen, 1997). Thus, it is important to show that the high congruence coefficients are not an artifact of the rotation procedure. Following the procedures of previous simulation work (Paunonen, 1997), we generated 1000 bootstrapped matrices for each world region, derived from the region's original matrix. To bootstrap each matrix, we shuffled the original matrix of the mean trait ratings (columns) of all faces (rows). To make the test more stringent, rather than randomly shuffling values all across the matrix (as in e.g., Paunonen, 1997) or randomly generating values (as in e.g., Korth, 1978), we shuffled the values while keeping them within each of their original columns (i.e., trait judgment), leaving some original characteristics of the matrix intact. The mean rating matrix of the Todorov Lab data was left

Table 4 Factor congruence scores between Todorov Lab data of judgments of faces (Oh, Dotsch, Porter, et al., 2019; Oosterhof & Todorov, 2008) and PSA replication data of judgments of a different set of faces in 11 world regions (Jones et al., 2020).

	PCA		EFA	
World region	**PC 1**	**PC 2**	**Factor 1**	**Factor 2**
Africa	0.992	0.972	0.994	0.968
Asia	0.986	0.885	0.986	0.887
Australia & New Zealand	0.992	0.980	0.992	0.983
Central America & Mexico	0.995	0.961	0.993	0.969
Eastern Europe	0.981	0.989	0.983	0.983
Middle East	0.975	0.916	0.982	−0.884[a]
Scandinavia	0.989	0.986	0.990	0.985
South America	0.988	0.966	0.989	0.970
United Kingdom	0.990	0.966	0.992	0.968
USA & Canada	0.987	0.977	0.990	0.976
Western Europe	0.991	0.969	0.993	0.966

[a]The congruence coefficient for the Middle East is negative, because the signs of the loadings of the judgments are reversed. However, the judgments with highest loadings remain dominance, aggressiveness, and meanness as in the Todorov lab data (Table 3). EFA results are also highly dependent on the rotation procedure. With a different oblique rotation—promax—this congruence coefficient is 0.909; and the range of congruence coefficients is from 0.983 to 0.995 for Factor 1 and from 0.889 to 0.985 for Factor 2. Congruence scores for both PCA and EFA are computed after a Procrustes rotation of the data from each world region to align with the Todorov Lab data.

unaltered. We computed the congruence coefficients between each of the 1000 bootstrapped matrices and the Todorov Lab matrix for each component/factor after a Procrustes rotation. We then constructed a distribution of the congruence coefficients, taking the absolute values of the coefficients. Note that the latter procedure creates an upward shift in the coefficients and, thus, makes it more difficult to reject the hypothesis that the observed congruence coefficients are an artifact of the Procrustes rotation.

Fig. 4 shows the distributions of the simulated congruence coefficients for the first two principal components based on the PCA simulations. In all 11 regions, the observed congruence coefficients for both the first and the second PCs are much higher than the simulated coefficients (Africa: $M \pm SD = 0.32 \pm 0.18$ and 0.26 ± 0.18, Asia: 0.32 ± 0.19 and 0.26 ± 0.17, Australia and New Zealand: 0.32 ± 0.19 and 0.26 ± 0.18, Central America

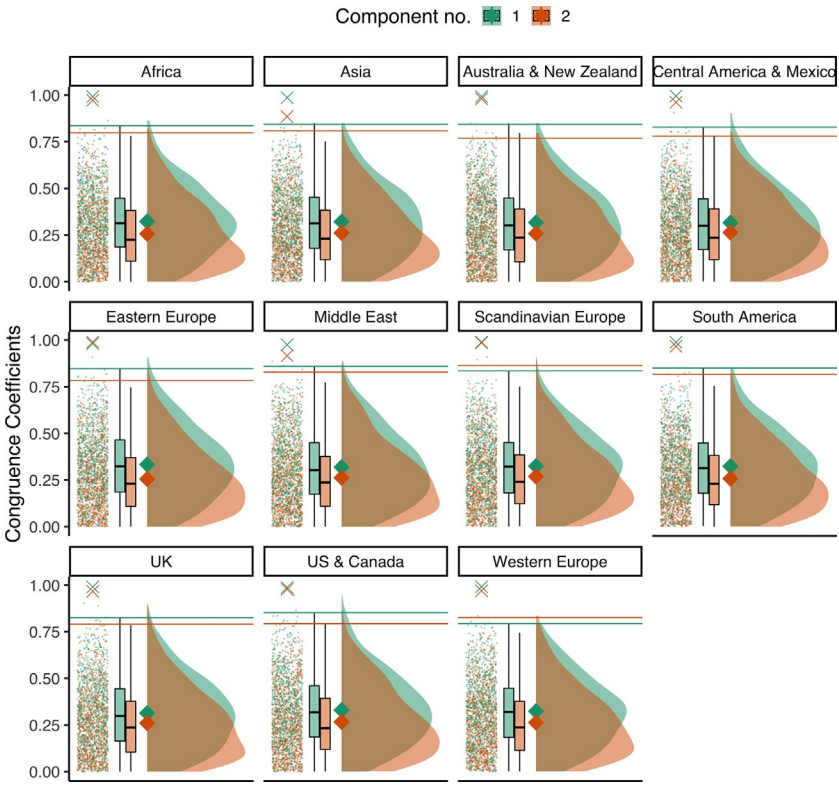

Fig. 4 Simulated and observed congruence coefficients from principal component analyses after a Procrustes rotation. Simulated congruence coefficients were computed between the Todorov Lab data and 1000 bootstrapped samples from every region (dots = individual datapoints; black lines in the boxplot = median; diamonds on the violin = mean). The simulated coefficients (dots) are plotted against the actual, observed congruence coefficients (the "x" above dots). The observed congruence coefficients are higher than the simulated ones (horizontal line = the 99.9th percentile of simulated coefficients), suggesting that the high level of congruence in each region is extremely unlikely to be an artifact of the Procrustes rotation.

and Mexico: 0.32 ± 0.18 and 0.26 ± 0.17, Eastern Europe: 0.33 ± 0.19 and 0.26 ± 0.18, Middle East: 0.32 ± 0.19 and 0.26 ± 0.18, Scandinavian Europe: 0.32 ± 0.18 and 0.27 ± 0.18, South America: 0.32 ± 0.18 and 0.26 ± 0.17, UK: 0.31 ± 0.18 and 0.26 ± 0.18, US and Canada: 0.33 ± 0.19 and 0.27 ± 0.18, Western Europe: 0.32 ± 0.18 and 0.26 ± 0.18). In fact, none of the 1000 simulated coefficients exceeds the observed coefficients. In terms of statistical significance, this corresponds to $P < 0.001$ for every region, making it extremely unlikely that the high values of the observed congruence

coefficients are an artifact of the Procrustes rotation. Again, in light of the similarity of the correlational structures of the data (Fig. 3), this should not be surprising.

The structure of judgments derived from the EFA is also remarkably consistent across cultures (Table 4). For the first factor, the range of the factor congruence coefficients is from 0.982 to 0.994. That is, this factor would be considered structurally identical in all world regions. For the second factor, the range is from 0.884 to 0.985, though the sign of the coefficient for the Middle East is negative. However, the interpretation of this factor in the Middle East remains the same (see note to Table 4), and with a different oblique rotation, the sign is positive (we discuss rotation procedures at the end of the section). As in the case of the PCA results, this factor would be considered identical in nine world regions and very similar in the remaining two regions (Asia and the Middle East).

Fig. 5 shows the distributions of the simulated congruence coefficients for the first two factors based on the EFA simulations. As in the case of the PCA, in all 11 regions, the observed congruence coefficients for both the first and the second factors are much higher than the simulated coefficients (Africa: $M \pm SD = 0.33 \pm 0.17$ and 0.26 ± 0.18, Asia: 0.33 ± 0.18 and 0.27 ± 0.19, Australia and New Zealand: 0.32 ± 0.17 and 0.27 ± 0.18, Central America and Mexico: 0.32 ± 0.17 and 0.27 ± 0.18, Eastern Europe: 0.34 ± 0.17 and 0.27 ± 0.18, Middle East: 0.33 ± 0.17 and 0.27 ± 0.19, Scandinavian Europe: 0.33 ± 0.17 and 0.27 ± 0.18, South America: 0.33 ± 0.17 and 0.28 ± 0.19, UK: 0.32 ± 0.17 and 0.28 ± 0.18, US and Canada: 0.33 ± 0.17 and 0.28 ± 0.19, Western Europe: 0.33 ± 0.17 and 0.27 ± 0.18). Again, the observed coefficients are higher than every single simulated coefficient, making it extremely unlikely that the high values of the observed congruence coefficients are an artifact of the Procrustes rotation. In sum, the simulation results show that the stable structural similarity across regions is extremely unlikely to be due to chance.

It should be noted that in the preliminary report of the PSA replication project (Jones et al., 2020), the factor congruence coefficients were computed without a proper Procrustes transformation to align the loading matrices. Nevertheless, the qualitative interpretation of the findings for the PCA remains the same. However, as explained in the previous section, this transformation is crucial for EFA results because of the indeterminacies of factor solutions. For example, without the transformation, the congruence coefficients for Asia are 0.839 and 0.513 for the first and second factor, respectively, which are substantially lower than the correct values. Essentially, not aligning

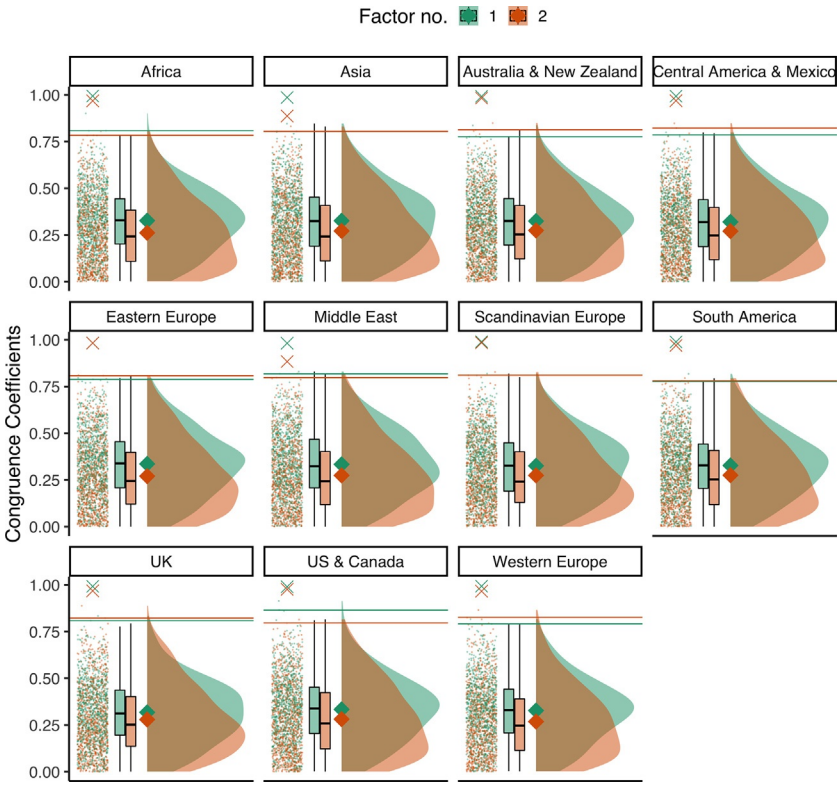

Fig. 5 Simulated and observed congruence coefficients from two-factor exploratory factor analyses after a Procrustes rotation. Simulated congruence coefficients were computed between the Todorov Lab data and 1000 bootstrapped samples from every region (dots = individual datapoints; black lines in the boxplot = median; diamonds on the violin = mean). The simulated coefficients (dots) are plotted against the actual, observed congruence coefficients (the "x" above dots). The observed congruence coefficients are higher than the simulated ones (horizontal line = the 99.9th percentile of simulated coefficients), suggesting that the high level of congruence in each region is extremely unlikely to be an artifact of the Procrustes rotation.

the loading matrices before computing similarity coefficients amounts to relying on the vagaries of statistical noise to make substantive inferences.

As explained above, we preferred to work with a 2D solution because of its simplicity and interpretability. Still, we can also specify a 3D solution and see how that affects the similarity of the structure of judgments in our dataset with the datasets from the world regions. For the PCA, the results remain the same for the first two interpretable PCs. The range of congruence coefficients is from 0.978 to 0.996 for the first PC and from 0.918 to 0.990 for

the second PC. The results are not as strong for the third PC, where the coefficients range from 0.812 to 0.911 in magnitude.

Unlike the third PC, the third factor in the EFA is interpretable. It contrasts judgments of unhappiness, weirdness, and meanness with judgments of emotional stability and sociability. This factor might be interpreted as emotional adjustment and it is strongly positively correlated with the first factor (0.66), interpretable as valence, and weakly negatively correlated with the second (−0.19), interpretable as power (the correlation between the first two factors is −0.11). In contrast to the PCA results, the results for the EFA change. Although the congruence coefficients range from 0.931 to 0.987 for the first factor, they range from 0.653 to 0.984 for the second factor and are below 0.85 for five regions (Africa, Asia, Australia and New Zealand, UK, US and Canada) with no apparent interpretation of the grouping of regions. For the third factor, the congruence coefficients are highly varied and are above 0.85 for only three regions (Eastern Europe, South America, Western Europe).

The instability of the congruence coefficients for the second factor is disconcerting, but their magnitude is heavily influenced by the rotation procedure (i.e., oblimin, as used by Jones et al., 2020). With another oblique rotation—promax—in a three-factor solution, the second factor's congruence coefficients range from 0.825 to 0.989, with only 1 coefficient below 0.85 (Western Europe); and with a simplimax rotation, the second factor's coefficients range from 0.916 to 0.992. Similarly, these coefficients are very high for orthogonal rotations (e.g., using quartimax, the range is from 0.910 to 0.988). Irrespective of the rotation procedure, the congruence coefficients are invariably high for the first factor (>0.95). Given that there are no mathematical reasons favoring one rotation procedure over another, one needs to have a strong theoretical rationale justifying the use of a specific procedure.

In sum, when a 2D structure is specified a priori, the simplistic 2D model, which we derived from judgments of US participants of white faces, seems to hold remarkably well in all world regions, in which data have been collected. When a more complex 3D structure is specified, the results for the first dimension—valence—are highly stable. Irrespective of the analytic (PCA vs. EFA) and rotation procedures (within EFA framework), this first dimension of face evaluation would be considered the same across world regions. The second dimension—power—is stable for the PCA results, but not for the EFA results, because the congruence coefficients change as a function of the rotation procedure. The least consistent results are obtained

when an oblimin oblique rotation is used. Interestingly, in this analysis the third factor is most interpretable, but its high correlation with the first factor (0.66) defeats the purpose of identifying independent dimensions of face evaluation.

1.3 Misunderstandings of structure models and cross-cultural differences

Recently we came across an anonymous review that claimed that dimensional structure models of judgments are clearly false, because the reviewer could not possibly imagine how the same face could be evaluated in the same way in two different regions of the US, let alone in different parts of the world. We do not take this as a failure of imagination, but as a failure to grasp what these models do. These models are about the correlational structure of judgments. That is, they tell us what kinds of evaluations are similar to each other. They do not tell us the exact value of evaluation of specific faces or groups of faces. Consider judgments of male and female faces. The correlational structure of judgments of male faces is similar to the correlational structure of judgments of female faces (but see Oh, Dotsch, Porter, et al., 2019 for subtle differences), but male and female faces are evaluated differently on most judgments. On average, female faces are evaluated as more trustworthy and attractive than male faces, but less confident and dominant. Thus, within the same structure, there might be large differences in evaluation as a function of the type of faces and individual characteristics of the raters (Martinez et al., 2020). By extension, one would expect differences as a function of the world region in which the ratings originate to the extent that these regions are populated with different raters who have different cultural beliefs.

Fig. 6 illustrates these points. On average, the two faces in Fig. 6A were evaluated as more positive and powerful than were other faces (note their location in the first quadrant of the valence/power judgment space). However, this general tendency is not uniform across regions. The variation in valence evaluation is extremely large for the White male face, with positive evaluations in some regions (e.g., Middle East) and negative in other regions (e.g., US and Canada). Similarly, the variation in power evaluation for the Black female face is large, with high evaluations in some regions (e.g., UK) and low in others (e.g., Africa). Finally, the Black female face is evaluated more positively than the White male face in most regions, but not in Central America and Mexico, South America, the Middle East, and Asia. We can also compare the valence evaluation of male and female faces

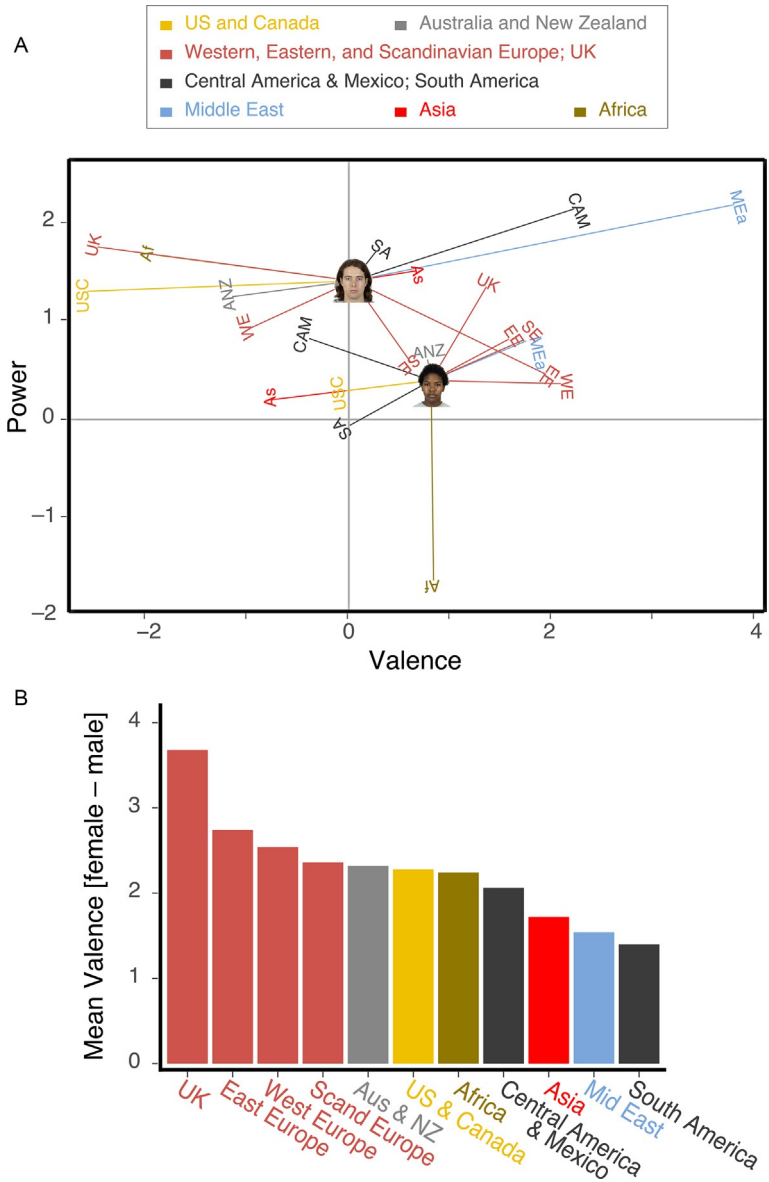

Fig. 6 Variations in valence and power evaluations of faces as a function of geographical region. (A) Valence and power scores of two sample faces, derived from ratings from 11 regions. These scores were computed by weighting the 13 trait judgments in each region by the coefficients of the first two PCs derived from a PCA on the mean face ratings aggregated across regions. This computation preserves differences that are specific for each region. (B) The mean difference in valence valuation of female and male faces in 11 regions. (Af = Africa, As = Asia, ANZ = Australia and New Zealand, CAM = Central America and Mexico, EE = Eastern Europe, MEa = Middle East, SA = South America, SE = Scandinavian Europe, USC = US and Canada, WE = Western Europe).

(Fig. 6B). While in all regions female faces are evaluated more positively, the magnitude of the difference varies across regions. The biggest difference is in European regions and the smallest in South America, the Middle East, and Asia.

Even within the same dimensional structure of judgments, there might be meaningful cultural differences. As should be obvious from Fig. 3, the correlational structure of judgments is highly similar across regions. This is made explicit in Fig. 7A. The pairwise correlations between regions are very high, with the lowest correlation being $\rho = 0.89$ (between the Middle East and Africa; a non-parametric Spearman correlation).

To assess the relative dissimilarity across regions, we submitted the pairwise Euclidean distances between all pairs of world regions and the world average to Kruskal's non-metric Multidimensional scaling (MDS) (Kruskal, 1964), a dimensionality reduction technique for an ordinal dataset. For visualization and interpretability, we reduced the data to two dimensions (Fig. 7B). In an MDS solution, points (regions in this case) that are closer together are more alike than those farther apart. Consistent with the similarity matrix (Fig. 7A), regions that can be grouped as similar (e.g., typically "Western," such as Western Europe, Scandinavia, the UK, and the US and Canada) are highly similar to each other, whereas regions that are

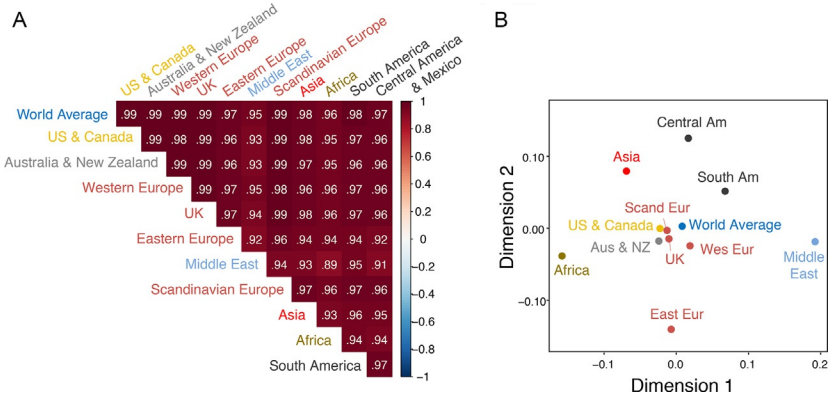

Fig. 7 Cross-regional similarities and differences in social judgments from faces. (A) Spearman rank correlational coefficients for all pairs of regions. These pairwise coefficients are computed from the vectors of Pearson correlations between judgments (see Fig. 3). (B) A non-metric multidimensional scaling visualizing relative dissimilarity across regions. *Data from Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., et al. (2020). To which world regions does the valence-dominance model of social perception apply? (Registered Report Stage 2).* PsyArXiv. *https://psyarxiv.com/n26dy*.

dissimilar from these regions (e.g., not typically "Western," such as Africa, Asia, the Middle East, and Central America) are distinct from these regions as well as from each other. We should emphasize that we are not making claims about the nature of cross-cultural differences here, but just pointing out that they could be found if this were the main interest of the researcher.

1.4 Limitations of dimensional structure models

Dimensional structure models of judgments reduce a large number of judgments to a few summary dimensions. Although these models are simple and elegant, they are largely descriptive and may not be very powerful when it comes to specific predictions (Todorov, 2009). In specific contexts (e.g., voting), judgments of attributes important for the context (e.g., perceived competence) are going to be more predictive of the final decision than judgments extracted from summary dimensions (Hall, Goren, Chaiken, & Todorov, 2009; Todorov et al., 2005). Hence, if the main objective is a prediction of a specific decision or a behavior, the best course of action is to first identify the attributes that matter in the particular context.

The power of dimensional structure models is that they provide a framework for thinking about judgments. To the extent that the same summary dimensions emerge across domains and cultures (Cuddy et al., 2009; Fiske & Cuddy, 2006; McCrae, 2002; Osgood, 1952; Osgood, Suci, & Tannenbaum, 1957; Saucier, 2003; Wiggins, 1979; Wiggins & Pincus, 1992), one can argue that these dimensions are universal and are in the service of the same social functions (e.g., figuring out the intentions and capabilities of others in the absence of good information; Todorov, 2017). An alternative interpretation is that the dimensions simply reflect semantic features of the language and/or implicit personality theories (Schneider, 1973; Stolier, Hehman, Keller, Walker, & Freeman, 2018). However, these interpretations are not necessarily mutually exclusive to the extent that common semantic features are grounded in reality. Moreover, as we show in the second part of this chapter, both general dimensions of judgments and specific judgments map onto distinct configurations of physical facial features.

The analyses in Section 1.2 suggest that the two summary dimensions we identified 12 years ago generalize across cultures. However, it would be a mistake to argue that these are the "ultimate" dimensions. The more modest inference is that given the same set of 13 judgments, one would almost always observe the same two dimensions, interpretable as valence and power, respectively. The most important determinant of the results of any dimensionality

reduction analysis is the input to this analysis. A different and a larger set of traits will result in a different dimensional solution. The only safe bet is that the first dimension will be about valence, something that we have known since the seminal work of Charles Osgood in the 1950s (Osgood, 1952; Osgood et al., 1957).

The set of faces is also important for the observed results. A good example is the work of Clare Sutherland and colleagues (Sutherland et al., 2013). Using a much larger and diverse set of faces, varying in age, they found that in addition to the dimensions, outlined in our work, a third dimension emerged: youthfulness/beauty. We were actually quite surprised to find the excellent correspondence between our own data and the PSA replication project data, given the differences in faces. However, although the faces used by the latter were more diverse than the faces we used, their variation, including on age, is much more limited than the variation of faces in the world. Finally, the content of the judgments and the nature of the faces can interact. When judging kids' faces, the second dimension is not about power, but about shyness (Collova, Sutherland, & Rhodes, 2019).

2. Revealing the perceptual basis of social judgments from faces

The first part of this chapter dealt with the first question that motivated our research: what is the structure of social judgments from faces? The second part deals with the second question: what is the perceptual basis of these judgments? As pointed out in the introduction, while the first question is informative about the second question, these two questions are conceptually and empirically independent. That is, one can try to figure out the perceptual basis of any judgment irrespective of how this particular judgment relates to any other judgment. Similarly, one can look for a structure of social judgments irrespective of the perceptual determinants of these judgments. However, findings about the structure of social judgments inform what judgments should be prioritized in terms of studying their perceptual basis. Given the 2D model of social judgments, it was natural to first build computational models of trustworthiness and dominance judgments, as those judgments best approximated the first two PCs, respectively (Oosterhof & Todorov, 2008; Todorov et al., 2008). More importantly, one can try to establish the nature of correspondence between these two spaces: social judgment space and physical face space.

In Section 2.1, we describe how a physical face space (Fig. 1B) can be built from a statistical analysis of variations of real faces. In Section 2.2, we show how social judgments (Fig. 1A) can be modeled and visualized in a data-driven way, using the physical face space (Fig. 1B). In Section 2.3, we show how this modeling approach can be used to control for shared variance between judgments (e.g., competence and attractiveness) and identify configurations of features that are specific to the judgment of interest (e.g., competence). We also show how the approach can be extended to measures other than explicit judgments, including neural measures. In Section 2.4, using a canonical correlation analysis, we outline an alternative approach that maps linear combinations of multiple social judgments rather than a single judgment (e.g., competence) to the physical face space. Finally, in Section 2.5, we discuss the limitations of data-driven computational models.

2.1 Tools to model social judgments from faces

Standard theory-driven methods to study the perceptual basis of judgments from faces have inherent limitations (Adolphs, Nummenmaa, Todorov, & Haxby, 2016; Todorov, Dotsch, Wigboldus, & Said, 2011), the most important being that the space of possible hypotheses is infinitely large. In the standard approach, one experimentally manipulates features (e.g., the degree of smiling or the distance between the eyebrows and the eyes) and observes how changes in these features affect specific judgments (e.g., friendliness or trustworthiness). The problem is that 10 binary features generate more than 1000 combinations and 20 binary features generate more than 1 million. To use a historical example, in the 19th century the British painter Alexander Cozens experimented with drawings of simple face profiles to discover the principles of human beauty (Todorov, 2017). He worked with relatively few variations: forehead (4 types), nose (12), mouth (16), chin (2), eyebrow (12), and eyes (16). Were Cozens a proper experimental psychologist, he should have created 294,912 different face profile drawings and have them judged by participants. Even in the age of online experiments, this would have been a hugely complicated experimental design to administer. There are even deeper problems than the rapid proliferation of feature combinations. Most features are neither binary nor discrete. Moreover, we do not even know what constitutes a proper feature: is it a mouth, a lip, a right corner of the lip, a pixel in a 2D space or a voxel in a 3D space? Finally, features deemed unimportant by the experimenter, although they could be critical for judgments, cannot be revealed by standard methods, because they are simply not manipulated

(Dotsch & Todorov, 2011). As an example, most people probably think that the eyes are more important than the eyebrows for face recognition. In fact, the opposite is true (Sadr, Jarudi, & Sinha, 2003).

Given the limitations of the standard methods, we have developed computational data-driven methods, which do not rely on prior theoretical hunches to discover the perceptual basis of judgments from faces. The computational foundation of these methods was provided by the work of Blanz and Vetter (1999), who developed a morphable statistical model of face representation derived from the analysis of 3D laser scans of real faces. We used a commercial implementation of their methods, FaceGen (Singular Inversions, 2005). FaceGen was created from 3D laser scans of the faces of about 300 people. To create a statistical model of face shape, a mesh with fixed topology was superimposed on the shape of each face and a PCA was conducted on the vertices of this mesh. The PCA extracted 50 PCs that define the shape of any face.

Fig. 8 illustrates the shape variations that define four of the 50 shape PCs. We choose the first two PCs that account for more variance than any other pairs of PCs and two other PCs that account for much less shape variance: PC 10 and PC 20. As in any PCA, each component accounts for less variance than the preceding component. The first PC appears to be related to the overall wideness of the face, but it should be noted that all face features are affected: nose, jaw, eyes, placements of the eyes, and so on (Fig. 8A). The second PC appears to be related to the elongation of the face, but it also changes the appearance of the chin, the forehead, and the eyes. The changes captured by the 10th PC are much less pronounced than the changes captured by the first 2 PCs, because the former accounts for less shape variance than the latter. The 10th PC is primarily related to changes of the nose and mouth, but also affects the shape of the eyes (Fig. 8B). Similarly, the 20th PC is related to changes of the internal features, but the changes are most pronounced around the eyes.

Two things should be noted about the shape PCs. First, each PC captures a set of holistic shape changes: variations are not limited to a specific feature. Second, the PCs need not have psychological meanings. They are statistically extracted from the shape variation of real human faces. What is important is that they can describe the shape of any face. To illustrate, the face in the first quadrant of Fig. 8A is 1 SD on the first PC and 2 SD on the second PC. The face in the second quadrant of Fig. 8B is 2 SD on the 10th PC and −2 SD on the 20th PC. The PCs define the space of face shapes, and each face is a vector, which is a linear combination of the underlying PCs, in this space (Fig. 8C).

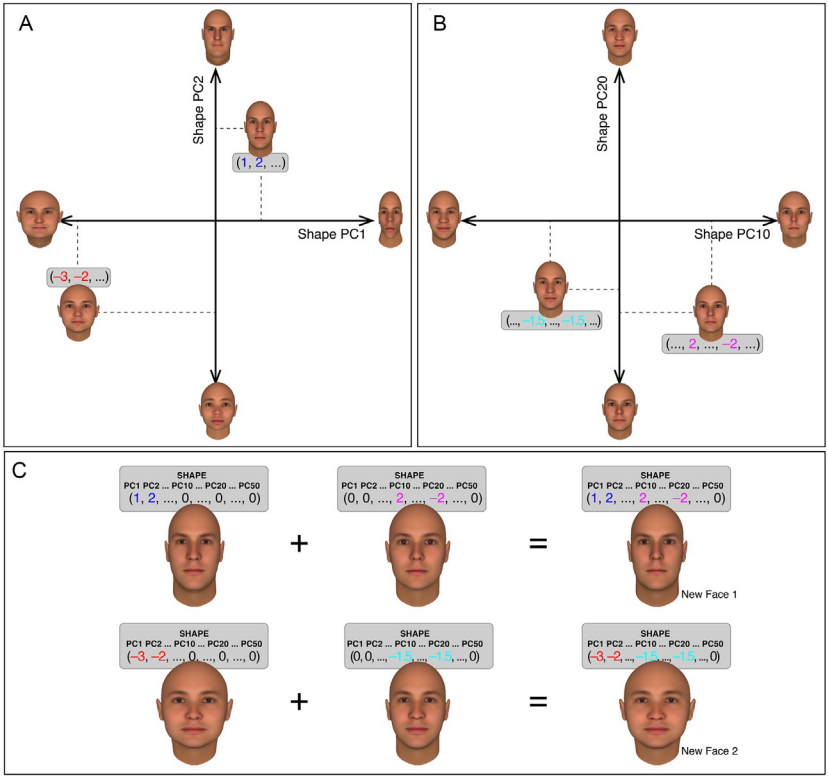

Fig. 8 Changes in facial appearances across four shape dimensions extracted from a statistical analysis of variation of real faces. (A) The first two dimensions (PCs) of a 50D model of face shape. (B) The 10th and 20th dimensions (PCs) of a 50D model of face shape. These two dimensions account for much less shape variance than the first two dimensions. In both panels, the anchor faces on the *x*- and *y*-axes are set at 6 SD to illustrate the shape changes along the respective dimensions. Two exemplar faces are shown in each panel with differences only on the respective two dimensions (the coordinates on the remaining dimensions are set at 0). (C) Two novel faces created by combining the exemplar faces. Novel faces are created by simply adding the values (functioning as parameters) on the respective dimensions.

It should be noted that these 50 PCs are not accessible in the commercially available version of FaceGen. What this version offers is "user-friendly" translations of the underlying PCs into sets of features such as "nose bridge—short/long". Changing the latter in FaceGen also changes all other "nose" control features (total of 14) and almost all other control features such as "chin—forward/ backward". That is, these features are not independent from each other and cannot be manipulated independently. In contrast, the underlying PCs are orthogonal and any face is a linear combination of

the PCs. For example, Fig. 8C shows how we can create new faces that are a combination of the faces shown in Fig. 8A and B. This statistical representation of faces allows us to randomly sample an infinite number of faces whose shape is completely determined by the 50 PCs.

A similar 50D model was derived for the reflectance of faces (brightness, texture, and color variation on the surface map of the face) based on a PCA of the RGB values of the color map of the scanned faces. The reflectance of faces is just as important as their shape, and sometimes more important, for a range of face processing tasks: from social judgments (Oh, Dotsch, & Todorov, 2019; Torrance, Wincenciak, Hahn, DeBruine, & Jones, 2014) to perception of facial expressions (Sormaz, Young, & Andrews, 2016) to person recognition (Andrews, Baseler, Jenkins, Burton, & Young, 2016; Caharel, Jiang, Blanz, & Rossion, 2009; Lee & Perrett, 2000; O'Toole, Price, Vetter, Bartlett, & Blanz, 1999; Russell, Sinha, Biederman, & Nederhouser, 2006; Troje & Bülthoff, 1996). Fig. 9 shows the reflectance variations that define the first two PCs, as well as the 10th and the 20th PCs. The first PC appears to be related to the overall whiteness/darkness of the face, but also to eye color and shading around the eyes and mouth. The second PC appears to be related to the coloration of the face, ranging from skin-like color to purple, but also to the shading of the eyebrows and whiteness of the eyes. The 10th PC is related to the coloration in the middle of the face: eyes, eyebrows, and nose. The 20th PC is related to the coloration around the cheekbones, eyebrows, and the region above the lips. As in the case of the shape PCs, each reflectance PC captures a set of holistic reflectance changes and the PCs need not have psychological meanings. They are statistically extracted from the color variation of real human faces.

The final statistical model consists of 100 PCs or dimensions (50 shape +50 reflectance). With this statistical representation, we can build a model of any judgment, provided participants agree on this judgment. The next section outlines the modeling approach.

2.2 Data-driven computational models of social judgments

To create a computational model of a social judgment, the first step is to randomly generate a sample of faces from the statistical multidimensional space. As described above, each face is a vector in this space, uniquely defined by its coordinates on the 100 orthogonal dimensions. For most of the models described in this section, we used 300 randomly generated faces, but we have used as many as 4000 faces (Said & Todorov, 2011). The second step is to ask participants to judge the sample of randomly generated faces on

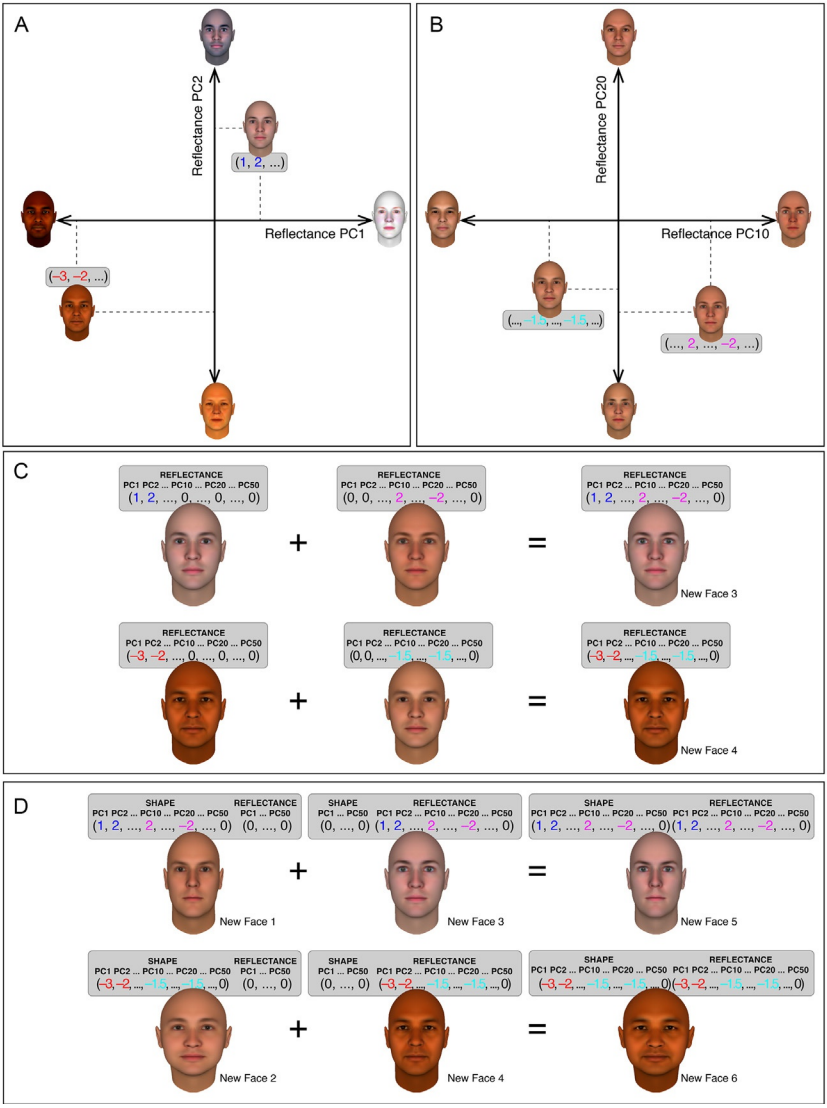

Fig. 9 Changes in facial appearances across four reflectance dimensions extracted from a statistical analysis of variation of real faces. (A) The first two dimensions (PCs) of a 50D model of face reflectance. (B) The 10th and 20th dimensions (PCs) of a 50D model of face reflectance. These two dimensions account for much less reflectance variance than the first two dimensions. In both panels, the anchor faces on the *x*- and *y*-axes are set at 6 SD to illustrate the reflectance changes along the respective dimensions. Two exemplar faces are shown in each panel with differences only on the respective two dimensions (the coordinates on the remaining dimensions are set at 0). (C) Two novel faces created by combining the exemplar faces in the first two panels varying only on reflectance. (D) Two novel faces created by combining the exemplar faces in Fig. 8A and B and Fig. 9A and B. Novel faces are created by simply adding the values (functioning as parameters) on the respective shape and reflectance dimensions.

the judgment of interest (e.g., trustworthiness). The third step is to establish that the participants' judgments are reliable.

This third step is extremely important, because one can build a model of any set of numbers, but if these numbers are not meaningful, the resulting model will be a model of noise. There are two facets of reliability. The first is the reliability of the ratings of the individual raters, most easily measured by their test-retest reliability. Notice that to measure the latter, participants must rate the faces twice. In our lab, it is a standard practice to have each participant rate a subset of the faces twice (e.g., Oh, Buck, & Todorov 2019). Raters with zero or negative test-retest reliability only contribute noise to the average judgment. Of course, if the main research interest is in finding the reliability of a particular judgment, the data from such raters should not be excluded from the analysis. The second facet of reliability is the inter-rater agreement. The latter varies widely across judgments and it is typically higher for judgments such as attractiveness (Oosterhof & Todorov, 2008, table S2). It should be noted that there are classes of stimuli where the test-retest reliability is high, but the inter-rater agreement is zero (Kurosu & Todorov, 2017; Martinez et al., 2020). In this particular case, it is possible to build a model of individual judgments: a model of aggregated judgments will be meaningless. All of the models described below are based on aggregated judgments, for which there is high inter-rater agreement (see Section 3 for the role of idiosyncratic contributions to judgments). Generally, the higher the inter-rater agreement, the better the resulting model is in terms of explaining the variance in judgments (Todorov & Oosterhof, 2011). Once it is established that the judgments are reliable, the last two steps are building the actual model and validating the model. We elaborate on these steps below.

As shown in Fig. 10, each face is a vector in a 100D space, though technically the shape and reflectance spaces are independent (see Todorov & Oosterhof, 2011), but additive (see Fig. 9D). Consider a matrix F (face matrix) consisting of 100 physical space coordinates (50 representing the shape and 50 representing the reflectance) of n randomly drawn faces, in which each column represents the coordinates of each of the n faces:

$$F = \begin{bmatrix} b_{1(1)} & \cdots & b_{n(1)} \\ \vdots & \ddots & \vdots \\ b_{1(100)} & \cdots & b_{n(100)} \end{bmatrix}$$

Consider now a vector consisting of the (centered) numeric human judgment of the n random faces (e.g., trustworthiness judgments ranging from

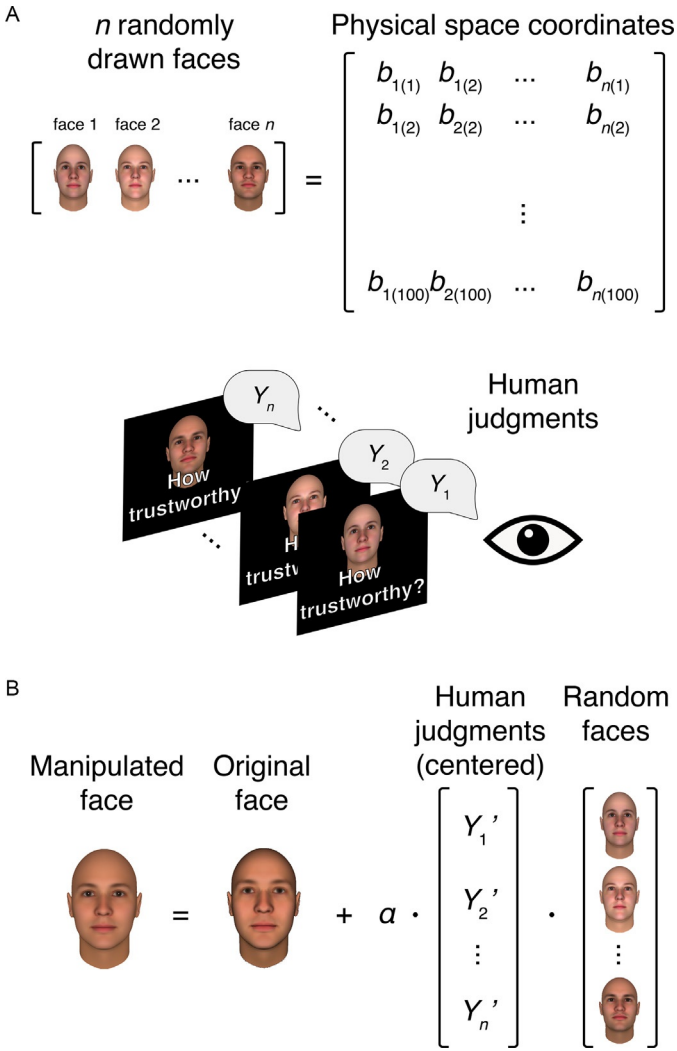

Fig. 10 Computing models of social judgments from faces. (A) A random sample of faces is drawn from the multi-dimensional physical face space. These faces are rated on a specific judgment. (B) By multiplying the average centered judgment with the coordinates of the randomly drawn faces, one can obtain a new vector optimal for changing faces on the judged dimension (typically, this vector is normalized): new faces can be changed with a tunable constant $\propto$.

-5 to 5), $\vec{j} = \left[Y'_1, Y'_2, \ldots, Y'_n\right]^T$ (Fig. 10A). By multiplying the judgments ($\vec{j}$, the judgment vector) by the coordinates of the randomly drawn faces (F, the face matrix), one can obtain a new vector in the face space optimal for changing the appearance of faces along the dimension of judgment,

$\vec{\Delta}$ (Blanz & Vetter, 1999). Specifically, the sums of the products of the judgments and the shape and reflectance parameters define the new vector of the judgment:

$$\vec{\Delta} = F \bullet \vec{j}$$

Divided by the norm or magnitude of the vector, the judgment vector is standardized: $\hat{\vec{\Delta}} = \vec{\Delta} / \|\vec{\Delta}\|$. This standardized vector, $\hat{\vec{\Delta}}$, constitutes the model of social judgment. One can now apply the model to any novel face to create a face that is meaningfully manipulated on the model representing the judgment, by any given unit (e.g., to create an individual face that appears +1 SD "trustworthy"; Fig. 10B). Consider a face, represented by a vector, $\vec{f}$, for example. One can manipulate the face to create a new face, $\vec{f}'$, by the unit of α (e.g., 1 SD) by simply:

$$\vec{f}' = \vec{f} + \alpha \bullet \hat{\vec{\Delta}}$$

This model computes the changes in the physical face space needed to change the perception of the face on the respective judgment. Fig. 11 illustrates this process for a model of judgments of trustworthiness and the changes along the first two shape and reflectance dimensions (components) in the physical face space. To evoke 1 SD unit change in the perceived trustworthiness of a face, the model changes 0.03 SD on the first shape component and −0.28 SD on the second component (Fig. 11A and B). We can infer that the second shape component is more important than the first one for judgments of trustworthiness. For reflectance, the first component is more important than the second one (Fig. 11A and C). We can also infer that the first two shape components are more important than the first two reflectance components for judgments of trustworthiness, because the length of the new shape vector is larger than the length of the new reflectance vector.

Fig. 11 and the previous paragraph are only meant to illustrate the mechanics of the models of the judgments, not an actual model. The actual models specify changes along all 100 components of the physical face space. Fig. 12 shows such models of judgments of extroversion, threat, and trustworthiness. These models capture the perceptual stereotypes of these personality traits. As the perceived extroversion of the faces increases (from left to right), the faces appear happier and their smiling increases.

A

Dimension	Coordinates			
Trustworthiness	–3.00	–1.00	1.00	3.00
Shape PC1	–0.09	–0.03	0.03	0.09
Shape PC2	0.84	0.28	–0.28	–0.84
Reflectance PC1	–0.33	–0.11	0.11	0.33
Reflectance PC2	–0.03	–0.01	0.01	0.03

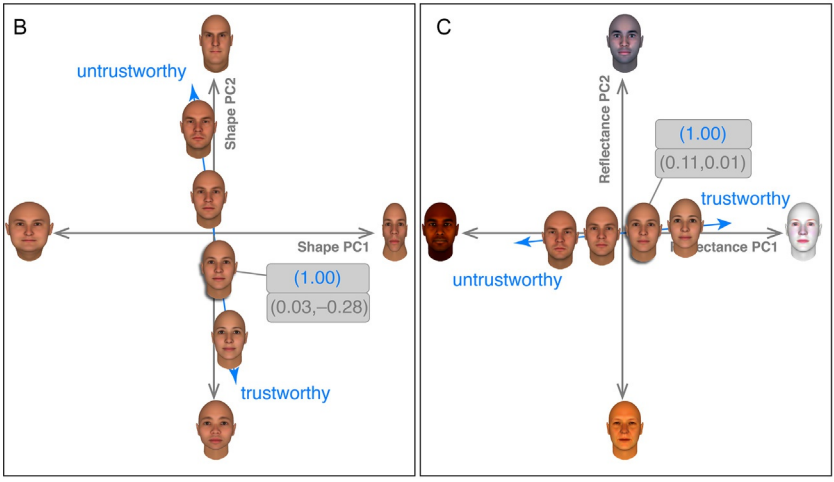

Fig. 11 Changes along the first two shape and reflectance dimensions to evoke a change in the perceived trustworthiness of faces. (A) The coordinates in SD units for corresponding changes in judgments and changes in shape and reflectance. (B) The direction of change along the first two shape components to increase or decrease the perceived trustworthiness of faces. ±1 SD and ±3 SD faces are shown along this new vector. (C) The direction of change along the first two reflectance components to increase or decrease the perceived trustworthiness of faces. ±1 SD and ±3 SD faces are shown along this new vector.

As the perceived threat increases, the faces appear more masculine. As the perceived trustworthiness increases, the faces appear more feminine and their smiling increases.

These models are data-driven, because in creating them we did not manipulate any features a priori. Our only "manipulation" was to randomly generate faces. That is, we allowed for the faces to vary randomly in the physical face space. Then we asked participants to judge these faces and based on these judgments extracted a new vector in the physical face space that is optimal in changing the shape and reflectance of faces to evoke corresponding

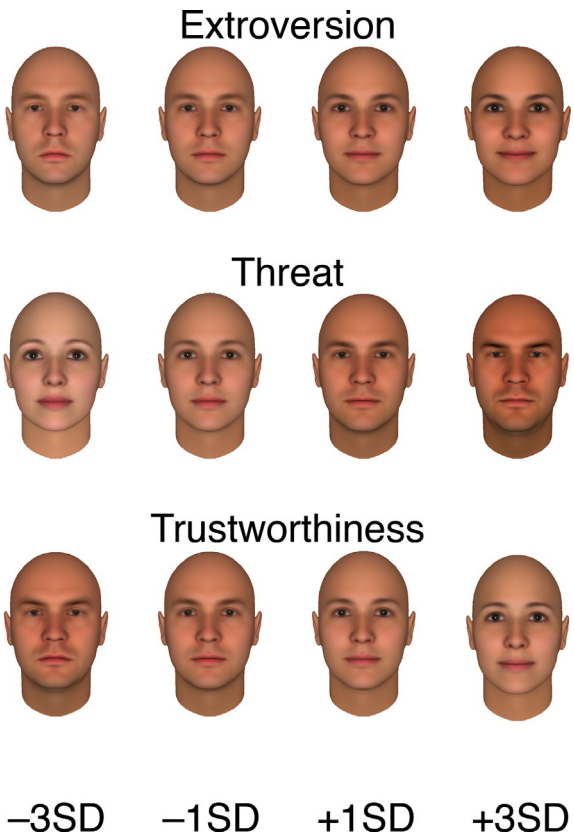

Fig. 12 Data-driven computational models of social judgments of extroversion, threat, and trustworthiness. The faces illustrate the changes in appearance that increase the judgments (from left to right) along the respective traits. The models were applied to the average face.

changes in judgments. Given this new vector model, we can project any existing face on the vector and increase or decrease its vector value to manipulate its appearance. We can think of these models as amplifiers of the signal in human judgments: amplifiers that identify a posteriori changes in appearance that are important for specific judgments. This data-driven method is a version of a reverse correlation approach (Todorov et al., 2011). In the standard theory-driven method, we manipulate features and observe changes in judgments. That is, we observe the correlation between manipulated features and judgments: a forward correlation. In the reverse correlation approach, we let the features randomly vary, observe a response and use this response to classify the random variation in features.

Having generated a model does not complete the process. Although typically the models show excellent face validity (no pun intended) as should be obvious to most readers from Fig. 12, they need to be validated (Todorov, Dotsch, Porter, Oosterhof, & Falvello, 2013). The typical validation procedure is to generate novel faces, manipulate them along the model, and ask a new group of participants to judge the manipulated faces. Participants' judgments are the ultimate criterion of whether the model is successful, because these are models of human perception. If the model is successful, participants' judgments should follow the model's predictions. Faces predicted by the model to appear more trustworthy, for example, should be judged as more trustworthy.

Within our lab, we have generated and validated many models: judgments of trustworthiness and dominance manipulated only as shape vectors (Oosterhof & Todorov, 2008); judgments of attractiveness (in Said & Todorov, 2011, we used a regression approach to model attractiveness in the statistical face space); judgments of attractiveness, competence, dominance, extroversion, likability, threat, and trustworthiness, manipulated as shape and reflectance vectors (Todorov et al., 2013); judgments of criminal appearance and remorse (Funk, Walker, & Todorov, 2016); judgments of physical strength and dominance (Toscano, Schubert, Dotsch, Falvello, & Todorov, 2016); judgments of competence controlling for judgments of attractiveness (Oh, Buck, et al., 2019); and judgments of trustworthiness controlling for judgments of attractiveness (Oh, Wedel, Labbree, & Todorov, in preparation). Outside our lab, using a similar approach, Mirella Walker and colleagues have generated and validated a number of models of other judgments: risk-seeking tendency and social skills (Walker, Jiang, Vetter, & Sczesny, 2011; Walker & Vetter, 2009), the Big Two in social cognition (e.g., warmth/communion, competence/agency, Abele & Wojciszke, 2007; Fiske, Cuddy, Glick, & Xu, 2002; Wiggins, 1979; Wiggins et al., 1989; Wojciszke, 1994), and the Big Five in personality psychology (e.g., agreeableness, concientiousness, extroversion, openness to experience, neuroticism, McCrae & Costa, 2008) (Walker & Vetter, 2016).

Based on our validated models, we have generated many face databases and have made those available for academic research. As of the end of April 2020, more than 4380 users—from over 900 institutions in 65 countries—have downloaded the face databases. These parametrically manipulated face stimuli have been used to study a wide range of questions in many fields in psychology, as well as outside of psychology. To list a few examples, in developmental psychology the stimuli have been used to study whether

infants are sensitive to facial "signals" of trustworthiness and dominance (Jessen & Grossmann, 2016), the developmental trajectory of social judgments from appearance in children (Charlesworth et al., 2019), and age-related differences between young and older adults (Cortes, Laukka, Ebner, & Fischer, 2019). In social psychology, behavioral economics, and political science, the stimuli have been used to study strategic interactions (Tingley, 2014), effects of appearance on economic transactions (Rezlescu, Duchaine, Olivola, & Chater, 2012), and voting preferences (Laustsen & Petersen, 2016).

2.3 Advantages of computational models of judgments

Having a model allows us to parametrically manipulate the appearance of any face. Moreover, as explained above, we can visualize the changes in appearance that are important for specific judgments. Although we worked with all 100 components of the physical face space to create models, a few components seem to capture most of the variation of judgments (Todorov & Oosterhof, 2011). It may be that the number of components needed scales up with the complexity of the perceptual task: from gender recognition to social judgment to person recognition (Kramer, Young, Day, & Burton, 2017). We have also found that the shape and reflectance components have similar (and largely additive) contributions to judgments (Oh, Dotsch, & Todorov, 2019). Finally, adding non-linear components to the models seems to add relatively little over and above linear components (Oh, Dotsch, & Todorov, 2019; Todorov & Oosterhof, 2011). However, consistent with recent work on the importance of face typicality for social judgments (Dotsch, Hassin, & Todorov, 2016; Sofer et al., 2017; Sofer, Dotsch, Wigboldus, & Todorov, 2015), computational work suggests that modeling typicality as a non-linear function substantially improves the ability of models to predict social judgments (Ryali, Goffin, Winkielman, & Yu, 2020; Ryali & Yu, 2018).

One of the most important features of our computational approach is that the models are in the same physical face space. This feature has three important implications: (a) the similarity of the models is immediately apparent; (b) we can control for shared variance between similar models; and (c) we can build models based on measures other than explicit judgments and relate these models to existing judgment models. The similarity of the models is indicated by the cosine of the angle between the vectors or the correlation of the respective models. Importantly, as shown in Fig. 13, the similarity of the models corresponds to the similarity of judgments.

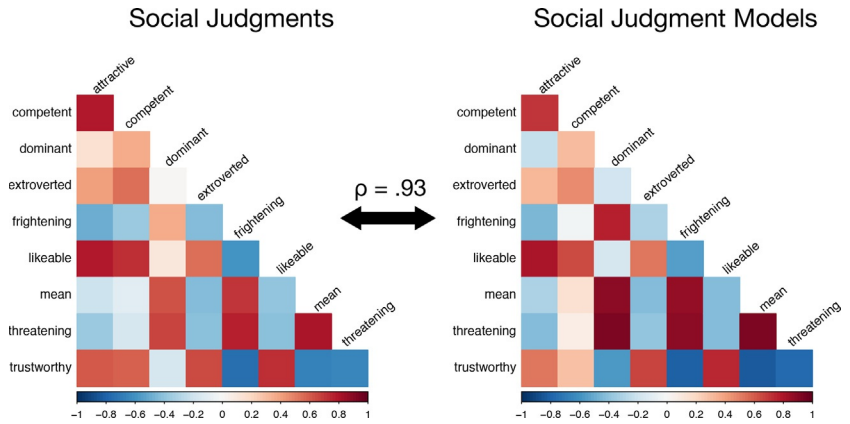

Fig. 13 Structural similarity between social judgments of faces and their computational models. Pairwise Pearson correlations of nine social judgments (of 300 faces) and correlations of nine models of these judgments (in a 100D physical face space). The correlational structures of judgments and models are highly similar to each other (measured using a non-parametric Spearman correlation). *Data from Todorov, A., & Oosterhof, N. (2011). Modeling social perception of faces.* IEEE Signal Processing Magazine, 28, *117–122.*

Fig. 13 also shows that two similar or highly correlated judgments (e.g., competence and attractiveness) will result in two highly similar models. The high inter-correlations between social judgments (see also part 1) pose both theoretical and experimental challenges. Consider a researcher interested in documenting the effects of perceived trustworthiness on cooperative exchanges or perceived competence on hiring decisions. Because perceived trustworthiness and perceived competence are highly correlated with attractiveness (Figs. 2C and 13), any observed effects of appearance can be attributed to attractiveness (e.g., the attractiveness halo effect; Dion, Berscheid, & Walster, 1972; Landy & Sigall, 1974; Thorndike, 1920). To avoid this ambiguity of interpretation, the researcher might resort to matching faces on attractiveness, but this strategy is generally suboptimal and can work only with a very large set of faces.

Within a computational framework, controlling for shared variance is straightforward. There are two procedures that one can use (Fig. 14). The first is to simply subtract from the model of interest (e.g., competence) the confounding model (e.g., attractiveness), effectively forcing the models to be negatively correlated. Note that the resulting models need to be validated. Faces manipulated on perceived competence after subtraction of attractiveness should be perceived as more competent but less attractive (Oh, Buck, et al., 2019). If one were to observe effects of appearance of

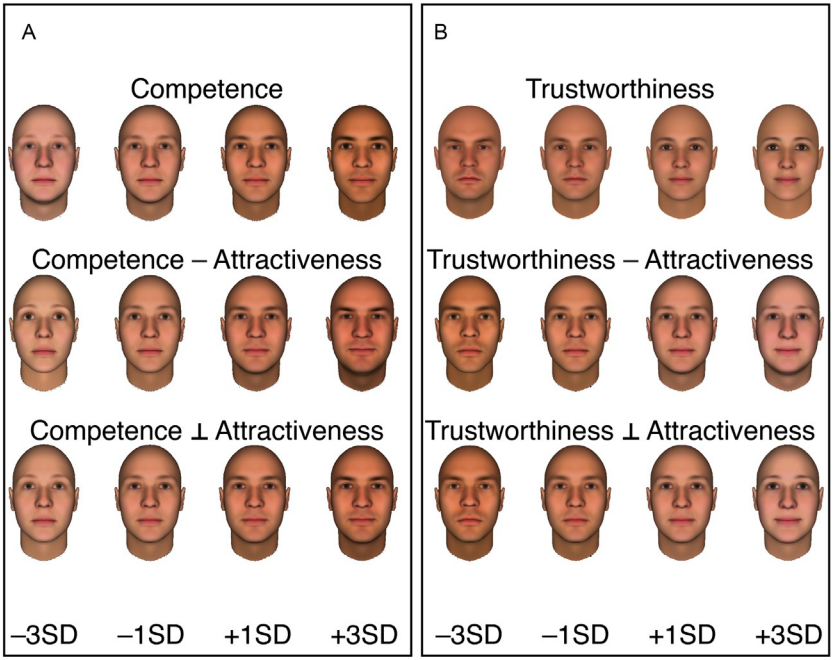

Fig. 14 Controlling for shared variance between highly correlated models of highly correlated judgments. (A) Models of judgments of competence. (B) Models of judgments of trustworthiness. The top row shows the original models, which are highly correlated with a model of attractiveness. The middle row shows the resulting models after subtracting the model of attractiveness. The bottom row shows models orthogonal to the model of attractiveness. *(Panel A) Adapted from Oh, D., Buck, E. A., & Todorov, A. (2019). Revealing hidden gender biases in competence impressions of faces.* Psychological Science, 30, *65–79. (Panel B) Adapted from Oh, D., Wedel, N., Labbree, B., & Todorov, A. (in preparation). Data-driven models of face-based trustworthiness judgments unconfounded by attractiveness.*

competence—manipulated by this particular model—on say hiring decisions, it should be clear that the effects cannot be attributed to attractiveness. The second procedure is to make the two models orthogonal. In principle, this procedure should guarantee that changes in judgments along the intended dimension (e.g., competence) are not accompanied by changes along the orthogonal dimension, but in practice judgments often change along the latter though to a much smaller extent. This is another reason to validate the resulting models.

Besides the experimental control for shared variance, these procedures of subtraction and orthogonalization can reveal important theoretical and

practical effects about the ingredients of the visual stereotypes of traits. In the case of competence, after controlling for attractiveness, the resulting competent-looking faces appear much more masculine, revealing the role of gender stereotypes in judgments of competence (Oh, Buck, et al., 2019). In the case of trustworthiness, after controlling for attractiveness, the resulting trustworthy-looking faces appear with much more positive expressions, revealing the role of emotion signals in judgments of trustworthiness (Oh et al., in preparation).

The third implication of the fact that the computational models are in the same physical face space is that new models based on measures other than explicit judgments can be immediately compared to existing judgment models. This type of comparison helps with the interpretability of the findings. These measures can range from response times to neural measures. Using a continuous flash suppression paradigm, in which visual input from one of the eyes is suppressed, Abir and colleagues measured the response times of detecting faces breaking out in consciousness (Abir, Sklar, Dotsch, Todorov, & Hassin, 2018). Based on these measures, with faster detection indicating prioritization of stimuli for conscious processing, they built a model of the face variation leading to faster responses. The resulting model was highly correlated with a model of judgments of dominance, but not a model of judgments of trustworthiness. Using fMRI, Cao and colleagues built models of face variation driving responses in the face-selective cortex (Cao, Li, Todorov, & Wang, 2020). They found that a model of this variation in the right fusiform face area was correlated with a model of judgments of trustworthiness, but only when participants were making trustworthiness judgments. In sum, the computational framework allows for linking models at different levels of face processing.

2.4 An alternative approach to mapping social judgments to physical face space

An alternative approach to modeling single social judgments such as trustworthiness is to model a combination of social judgments in the physical face space. We illustrate this approach here, using 9 social judgments of 300 randomly generated faces (Oosterhof & Todorov, 2008, table S6). In some ways, this approach comes closest to trying to directly relate the space of social judgments to the physical face space (Fig. 1), although its results may be less straightforward to interpret.

We use a canonical correlation analysis (Hotelling, 1936; Pedhazur, 1982; Sharma, 1996), which identifies linear combinations of two sets of variables

(e.g., judgments and physical features) that are maximally correlated with each other. In other words, one can identify linear combinations of judgments that are best predicted by configurations of physical features. The canonical correlation analysis is similar to PCA, as it identifies orthogonal linear combinations of variables. However, unlike PCA, where the objective is to explain maximum variance with as few PCs as possible, the objective is to maximize the correlation between the linear combination of variables. The number of linear combinations is equal to the number of the smaller subset of variables. In this particular case, this is the number of judgments. Also, unlike in PCA, it is possible to test for significance of each linear combination, called a canonical variate (CV) in this analysis. Table 5 shows the correlations between the 9 judgments and the first 5 CVs, which were significant.

As in PCA, the CVs need to be interpreted. The first CV is highly positively correlated with judgments of dominance, threat, meanness, and fright; and negatively correlated with judgments of trustworthiness. The second CV is highly positively correlated only with judgments of extroversion and to some extent with judgments of trustworthiness. The third, fourth,

Table 5 Correlations between the first five canonical variates (CV) and social judgments from faces.

Judgment	CV 1	CV 2	CV 3	CV 4	CV 5
Attractive	−0.10	−0.08	**0.49**	**0.51**	**0.67**
Competent	0.14	0.20	**0.53**	**0.63**	**0.38**
Dominant	**0.94**	0.17	0.04	0.23	0.17
Extroverted	−0.27	**0.80**	**0.34**	**0.36**	0.13
Frightening	**0.60**	−0.03	0.16	**−0.73**	−0.19
Likeable	−0.16	0.21	0.17	**0.58**	**0.63**
Mean	**0.80**	−0.25	0.21	−0.14	−0.32
Threatening	**0.87**	−0.04	−0.02	**−0.36**	0.13
Trustworthy	**−0.44**	**0.38**	−0.03	**0.48**	**0.52**
Canonical correlation	0.97	0.92	0.87	0.79	0.73

Correlations with a magnitude above 0.30 are in bold font.
The correlations here are equivalent to structure loadings in a PCA framework. They are different from the regression coefficients of the judgments, because the latter are highly correlated. These coefficients are equivalent to pattern loadings in EFA, when the factors are not orthogonal.

and fifth CVs are positively correlated with judgments of attractiveness and competence, but there are also meaningful differences. Judgments of extroversion are more highly correlated with the third and fourth CVs than with the fifth. Judgments of likeability and trustworthiness are more highly correlated with the fourth and fifth CVs than with the third. Finally, the fourth CV is highly negatively correlated with judgments of fright and to a weaker extent with threat.

The visualization of the CVs (Fig. 15) is consistent with the correlations described above, although these models have not been formally validated (see below). As the value of the faces increases on the first CV, they become more dominant and threatening looking. As noted in the previous section, once a new vector is built in the face space (e.g., CV 1), its similarity with existing vectors is immediately given. Consistent with the visual changes, the model of the first CV is significantly correlated with the models of judgments of dominance (0.48), threat (0.49), meanness (0.51), fright (0.45), and trustworthiness (−0.40). As the value of faces increases on the second CV, they become more extroverted looking. This model is only significantly correlated with the model of judgments of extroversion, although the correlation is weak (0.22). As the values of the faces increase on the third, fourth, and fifth CVs, they appear more competent and attractive. However, the changes on the third CV do not make the face appear more likeable and trustworthy, unlike the changes on the fourth and fifth CVs. Finally, the changes on the fourth CV also reduce perceptions of threat. Interestingly, the models of these three CVs are not significantly correlated with the models of any of the nine judgments, suggesting that the canonical correlation approach may reveal vectors in face space leading to the discovery of new information important for social judgments.

Table 5 also shows the correlation of each CV with the linear combinations of the shape and reflectance components of the physical face space (see Figs. 8–9). These components account for 94% of the variance in the first CV. It is instructive to compare the linear combinations of the judgments comprising the CVs with the linear combinations comprising the PCs extracted from a PCA of the judgments (Oosterhof & Todorov, 2008, table S6). First, these linear combinations are related but not the same. The correlation between the first CV and the first PC, interpreted as valence, is −0.56, whereas the correlation with the second PC, interpreted as power, is 0.78. Second, the physical features account for a larger percentage of variance in the first CV (94%) than in the first PC (72%) or the second PC (83%). This is a natural consequence of the canonical correlation analysis.

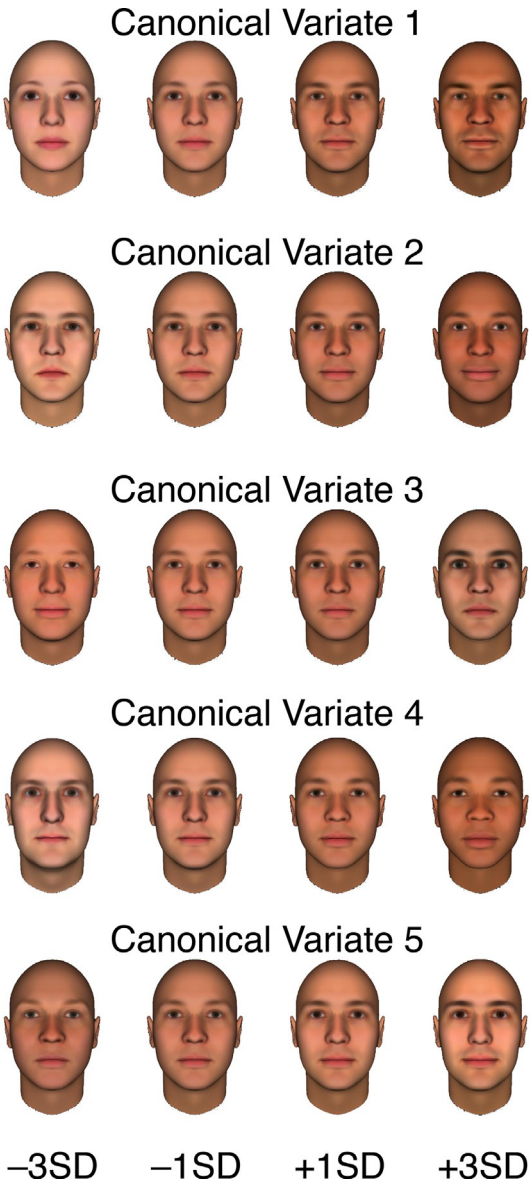

Fig. 15 Visualization of significant canonical variates. Each canonical variate is a linear combination of nine social judgments (see Table 5). The variates are orthogonal to each other, although the resulting models of the covariates are not. The models were applied to the average face.

Although the CVs and the PCs are linear combinations of the same judgments, the CVs are computed to maximize the correlation with the linear combination of face features.

The main advantage of this alternative approach of modeling social judgments is that it reveals what types of social judgments are best captured by physical features. The analysis above suggests that some judgments such as dominance and extroversion are more easily explainable by physical features. This approach may also reveal new directions in face space that can lead to interesting dissociations between judgments (e.g., compare CV 4 and CV 5 in Fig. 15). However, as a general rule, vectors that are a linear combination of judgments may be more difficult to interpret than vectors of specific judgments. The validation of such vectors of CVs is also not obvious, unless one validates the vectors on all judgments that form the CVs.

2.5 Limitations of data-driven models

One legitimate—though unsurprising—complaint about the faces manipulated by the models described above is that the faces are not realistic (Balas & Pacella, 2015; Crookes et al., 2015). In fact, the FaceGen faces lack distinctiveness and are therefore not as memorable and individualized as images of real faces (Balas & Pacella, 2015). This is the main reason that in all studies in our lab on learning affective associations with faces, we have used images of real faces (e.g., Falvello et al., 2015; Ferrari et al., 2020; Verosky et al., 2018; Verosky & Todorov, 2010, 2013). However, although the synthetic faces are not realistic, they can be morphed with images of real faces to enhance specific judgments (e.g., Jaeger, Todorov, Evans, & van Beest, 2020; Oh, Buck, et al., 2019; Oh, Dotsch, Porter, et al., 2019). Ulitmately, this criticism is tangential to the computational appoach. Moreover, there are existing methods of generating hyper-realistic faces (e.g., Karras et al., 2019; see Fig. 16).

The main limitation of data-driven methods is that features or variations that are important for the judgment or measure at hand, but are not present in the input space or faces, are not going to be discovered. Pragmatically, this comes down to the sample of faces used to create both models of physical face space and of judgments (Fig. 1). For example, the FaceGen model was created from laser scans of the faces of about 300 people, the majority of whom were white. Although the model can generate faces of different

Fig. 16 Perceived trustworthiness of hyper-realistic face images generated via a generative adversarial network (GAN) model. Participants rated the trustworthiness of 500 nonreal face images. Examples of faces perceived to be high (top), medium (middle), and low (bottom) on trustworthiness are shown here. Within each row, a left column corresponds to a higher perceived level of trustworthiness. The color of the box surrounding each face represents the mean trustworthiness rating of each face averaged across participants. The color bar represents the range of trustworthiness judgments across all 500 face images.

ethnicities, it is far from obvious that all the relevant variation needed to represent human diversity is present in the statistical face space.

The same considerations apply to models of judgments. The sample of faces used for the judgments would determine the resulting models. The FaceGen faces we used to build our models did not vary sufficiently on age. As a result, age cues are not prominent in our models. In contrast, Sutherland and colleagues, who used a large sample of real face images and created morphing continuums of social judgments, found that such cues are prominent in judgments of trustworthiness and dominance (Sutherland et al., 2013). Even in this work, however, only judgments of adult faces were collected. Not having kids' faces precludes the opportunity to observe variations important for social judgments (Berry & Zebrowitz McArthur, 1986; Montepare & Zebrowitz, 1998). The same considerations apply to variations of sex and ethnicity. The more representative the sample of faces of human diversity is, the more accurate the models will be.

3. Future directions

When we started the modeling work described above, computer science methods had not undergone the kinds of huge advancements we have seen in recent years owing to the application of deep neural networks. Thanks to such technologies, new methods for generation of face images are rapidly progressing (e.g., Karras et al., 2019). We are now capable of generating hyper-realistic faces that do not exist. Fig. 16 shows examples of such faces that have been rated on perceived trustworthiness. Several things should be noted. First, these faces do not exist in the world, though they might resemble people we know. Second, the cues identified in our model of trustworthiness (Fig. 12) are present in these images. More trustworthy faces are likelier to be female and to be smiling. Third, the most trustworthy faces are faces of kids, something that we have not observed in our prior studies, simply because we never used images of kids' faces. Of course, on the basis of her research, Leslie Zebrowitz would have predicted this finding (Berry & Zebrowitz McArthur, 1985, 1986; Montepare & Zebrowitz McArthur, 1986; Zebrowitz McArthur & Apatow, 1984). In the future, the computational models described in the second part of this chapter might be replaced by hyper-realistic models (Todorov, Uddenberg, Peterson, Griffiths, & Suchow, 2020). However, two things should be noted. First, the conceptual logic of model building remains the same. Second, we will be trading off hyper-realism of faces for less transparency and interpretability of the models.

All of the models presented here were based on judgments aggregated across participants. This aggregation masks the role of stable individual differences in social judgments from faces. While it is clear that these differences stem from observers' idiosyncratic statistical learning and personal knowledge (Dotsch et al., 2016; Stolier, Hehman, & Freeman, 2020; Sutherland et al., 2020; Verosky & Todorov, 2010, 2013; Xie, Flake, & Hehman, 2018), a quantitative assessment of such differences has been lacking. To measure the stable individual differences in face evaluation, one needs to have participants rate the same faces twice, a procedure that is almost never followed in practice. The result is studies that over-emphasize the role of consensus in judgments. However, even for the most appearance-based judgment—attractiveness—the contribution of consensus to judgments is 50% at best (Hönekopp, 2006; Martinez et al., 2020). Introducing repeated measurement allows for partitioning of the reliable variance into shared and idiosyncratic components.

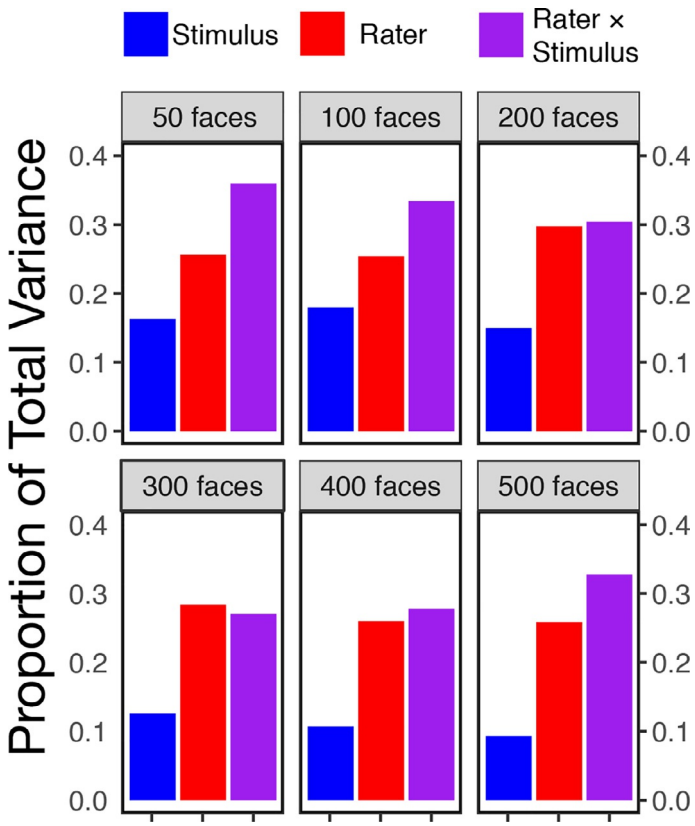

Fig. 17 Shared and idiosyncratic components of trustworthiness judgments of hyper-realistic faces created via a generative adversarial network (GAN). Participants were randomly assigned to rate 50, 100, 200, 300, 400, or 500 faces twice. The stimulus variance component captures differences between mean judgments of faces, and reflects consensus or agreement in judgments. The rater variance component captures differences between the mean judgments of raters. The rater × stimulus variance component captures interactions between raters and stimuli.

Fig. 17 shows the variance partitioning for judgments of the trustworthiness of hyper-realistic faces (Fig. 16). Irrespective of how many faces participants rated (ranging from 50 to 500), the stimulus contribution (e.g., some faces are more trustworthy-looking than others) is smaller than the contributions of the rater (e.g., the second author on average likes all faces more than the first author) and the rater by stimulus interaction (e.g., the first author likes face A more than face B, but the second author likes face B more than face A). While one can argue about the interpretation of the rater component (e.g., are differences in mean ratings meaningful in terms

of predicting behavior, or do they simply reflect biases in using the scale?), the rater by stimulus component reflects genuine idiosyncratic differences in judgments. To the extent that these differences contribute more to judgments than consensus, we are missing a big part of the picture by focusing on studying consensus. The next generation of studies should model both shared and idiosyncratic contributions to judgments.

4. Conclusions

The structure of social judgments from faces is highly stable. This is best seen from the similarity of the correlational structures across time (Fig. 2) and across cultures (Figs. 3 and 7A). In principle, this correlational structure would be compatible with multiple dimensional structure models depending on the type of analysis (e.g., PCA, EFA, canonical correlation analysis) and specific analytic procedures (e.g., rotation of dimensions). The dimensional solutions would also be affected by the content and number of the input judgments, and the variation of the faces. However, one thing should be clear: the final criterion is the interpretability of the dimensional structure model. There are no analytic solutions that guarantee the "right" model. For the set of traits we used 12 years ago, a 2D structure seems to account sufficiently well for the data.

Twelve years ago, we also proposed data-driven computational models of judgments, starting with models of judgments of trustworthiness, dominance, and threat. Since then we have validated many more models. These models, which do not rely on prior theoretical hunches of what face variations are important for social judgments, reveal the variations that drive specific judgments. The computational framework allows for precise parametric manipulation of the appearance of faces, for control of shared variance between judgments, and for relating models at different levels of face processing. Although the models we developed are going to be replaced by more sophisticated and hyper-realistic models, the conceptual logic of model building remains the same.

Acknowledgments

We thank Dan Albohn, Bertram Gawronski, Robin Gomila, Joel Martinez, and Stefan Uddenberg for comments on previous drafts of this manuscript. We thank Brandon Labbree for his help in data collection and generating images. We thank Ben Jones and Lisa DeBruine for initiating the cross-cultural replication of the structure of judgments from faces identified by Oosterhof and Todorov (2008). Data and scripts are available via the Open Science Framework at https://osf.io/ejh5r/.

References

Abdi, H. (2007). RV coefficient and congruence coefficient. In N. Salkind (Ed.), *Vol. 849. Encyclopedia of Measurement and Statistics:* (pp. 853–862). Sage.

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, *93*, 751–763.

Abir, Y., Sklar, A. Y., Dotsch, R., Todorov, A., & Hassin, R. R. (2018). The determinants of consciousness of human faces. *Nature Human Behaviour*, *2*, 194–199.

Adolphs, R., Nummenmaa, L., Todorov, A., & Haxby, J. V. (2016). Data-driven approaches in the investigation of social perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*, 20150367.

Ahler, D. J., Citrin, J., Dougal, M. C., & Lenz, G. S. (2017). Face value? Experimental evidence that candidate appearance influences electoral choice. *Political Behavior*, *39*, 77–102.

Andrews, T. J., Baseler, H., Jenkins, R., Burton, A. M., & Young, A. W. (2016). Contributions of feature shapes and surface cues to the recognition and neural representation of facial identity. *Cortex*, *83*, 280–291.

Antonakis, J., & Dalgas, O. (2009). Predicting elections: Child's play! *Science*, *323*, 1183.

Balas, B., & Pacella, J. (2015). Artificial faces are harder to remember. *Computers in Human Behavior*, *52*, 331–337.

Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences*, *104*, 17948–17953.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*, 269–278.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*, 1–68.

Berry, D. S., & Zebrowitz McArthur, L. (1985). Some components and consequences of a babyface. *Journal of Personality and Social Psychology*, *48*, 312–323.

Berry, D. S., & Zebrowitz McArthur, L. (1986). Perceiving character in faces: The impact of age-related craniofacial changes on social perception. *Psychological Bulletin*, *100*, 3–18.

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*, *19*, 187–194.

Caharel, S., Jiang, F., Blanz, V., & Rossion, B. (2009). Recognizing an individual face: 3D shape contributes earlier than 2D surface reflectance information. *NeuroImage*, *47*, 1809–1818.

Calder, A. J., Rhodes, G., Johnson, M., & Haxby, J. V. (Eds.). (2011). *Oxford handbook of face perception* Oxford University Press.

Cao, R., Li, X., Todorov, A., & Wang, S. (2020). A flexible neural representation of faces in the human brain. *Cerebral Cortex Communications*, *1*, 1–12.

Charlesworth, T. E. S., Hudson, S.-k. T. J., Cogsdill, E. J., Spelke, E. S., & Banaji, M. R. (2019). Children use targets' facial appearance to guide and predict social behavior. *Developmental Psychology*, *55*, 1400–1413.

Cogsdill, E. J., & Banaji, M. R. (2015). Face-trait inferences show robust child–adult agreement: Evidence from three types of faces. *Journal of Experimental Social Psychology*, *60*, 150–156.

Cogsdill, E. J., Todorov, A. T., Spelke, E. S., & Banaji, M. R. (2014). Inferring character from faces: A developmental study. *Psychological Science*, *25*, 1132–1139.

Collova, J. R., Sutherland, C. A., & Rhodes, G. (2019). Testing the functional basis of first impressions: Dimensions for children's faces are not the same as for adults' faces. *Journal of Personality and Social Psychology*, *117*, 900–924.

Colombatto, C., Uddenberg, S., & Scholl, B. J. (under review). The efficiency of demography in face perception.

Cortes, D. S., Laukka, P., Ebner, N. C., & Fischer, H. (2019). Age-related differences in evaluation of social attributes from computer-generated faces of varying intensity. *Psychology and Aging*, *34*, 686–697.

Crookes, K., Ewing, L., Gildenhuys, J. D., Kloth, N., Hayward, W. G., Oxner, M., et al. (2015). How well do computer-generated faces tap face expertise? *PLoS One*, *10*, e0141353.

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Leyens, J.-P., et al. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, *48*, 1–33.

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, *24*, 285–290.

Dotsch, R., Hassin, R. R., & Todorov, A. (2016). Statistical learning shapes face evaluation. *Nature Human Behaviour*, *1*, 1–6.

Dotsch, R., & Todorov, A. (2011). Reverse correlating social face perception. *Social Psychological and Personality Science*, *3*, 562–571.

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.

Dotsch, R., Wigboldus, D. H. J., & van Knippenberg, A. (2011). Biased allocation of faces to social categories. *Journal of Personality and Social Psychology*, *100*, 999–1014.

Falvello, V., Vinson, M., Ferrari, C., & Todorov, A. (2015). The robustness of learning about the trustworthiness of other people. *Social Cognition*, *33*, 368–386.

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A. T., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences*, *115*, E1690–E1697.

Ferrari, C., Oh, D., Labbree, B., & Todorov, A. (2020). Learning the affective value of people: More than affect-based mechanisms. *Acta Psychologica*, *203*, 103011.

Fischer, R., & Fontaine, J. R. J. (2011). Methods for investigating structural equivalence. In D. Matsumoto, & F. J. R. van de Vijver (Eds.), *Culture and psychology. Cross-cultural research methods in psychology* (pp. 179–215). Cambridge University Press.

Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology*, *10*, 1507.

Fiske, S. T., & Cuddy, A. J. (2006). Stereotype content across cultures as a function of group status. In *Social comparison and social psychology: Understanding cognition, intergroup relations, and culture* (pp. 249–263). Cambridge University Press.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*, 878–902.

Freeman, J. B., Rule, N. O., Adams, R. B., Jr., & Ambady, N. (2010). The neural basis of categorical face perception: Graded representations of face gender in fusiform and orbitofrontal cortices. *Cerebral Cortex*, *20*, 1314–1322.

Freeman, J. B., Stolier, R. M., & Brooks, J. A. (2020). Dynamic interactive theory as a domain-general account of social perception. In *Vol. 61. Advances in experimental social psychology* (pp. 237–287). Academic Press.

Funk, F., Walker, M., & Todorov, A. (2016). Modelling perceptions of criminality and remorse from faces using a data-driven computational approach. *Cognition and Emotion*, *31*, 1431–1443.

Goren, A., & Todorov, A. (2009). Two faces are better than one: Eliminating false trait associations with faces. *Social Cognition*, *27*, 222–248.

Hall, C. C., Goren, A., Chaiken, S., & Todorov, A. (2009). Shallow cues with deep effects: Trait judgments from faces and voting decisions. In E. Borgida, J. L. Sullivan, & C. M. Federico (Eds.), *The political psychology of democratic citizenship* (pp. 73–99). Oxford University Press.

Han, C., Wang, H., Hahn, A. C., Fisher, C. I., Kandrik, M., Fasolt, V., et al. (2018). Cultural differences in preferences for facial coloration. *Evolution and Human Behavior*, *39*, 154–159.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–83.

Hönekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 199–209.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*, 321–377.

Jaeger, B., Todorov, A., Evans, A., & van Beest, I. (2020). Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions. *Journal of Experimental Social Psychology*, *90*, 104004.

Jessen, S., & Grossmann, T. (2016). Neural and behavioral evidence for infants' sensitivity to the trustworthiness of faces. *Journal of Cognitive Neuroscience*, *28*, 1728–1736.

Johnson, K. L., Freeman, J. B., & Pauker, K. (2012). Race is gendered: How covarying phenotypes and stereotypes bias sex categorization. *Journal of Personality and Social Psychology*, *102*, 116–131.

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., et al. (2020). To which world regions does the valence-dominance model of social perception apply? (Registered Report Stage 2). *PsyArXiv*. https://psyarxiv.com/n26dy.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2019). Analyzing and improving the image quality of StyleGan. *arXiv*. https://arxiv.org/abs/1912.04958.

Korth, B. (1978). Superficiality and the dimensionality of sexism. *Applied Psychological Measurement*, *2*, 51–61.

Kramer, R. S. S., Young, A. W., Day, M. G., & Burton, A. M. (2017). Robust social categorization emerges from learning the identities of very few faces. *Psychological Review*, *124*, 115–129.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*, 1–27.

Kurosu, A., & Todorov, A. (2017). The shape of novel objects contributes to shared impressions. *Journal of Vision*, *17*, 1–20.

Landy, D., & Sigall, H. (1974). Beauty is talent: Task evaluation as a function of the performer's physical attractiveness. *Journal of Personality and Social Psychology*, *29*, 299–304.

Laustsen, L., & Petersen, M. B. (2016). Winning faces vary by ideology: How nonverbal source cues influence election and communication success in politics. *Political Communication*, *33*, 188–211.

Lawson, C., Lenz, G. S., Baker, A., & Myers, M. (2010). Looking like a winner: Candidate appearance and electoral success in new democracies. *World Politics*, *62*, 561–593.

Lee, K. J., & Perrett, D. I. (2000). Manipulation of colour and shape information and its consequence upon recognition and best-likeness judgments. *Perception*, *29*, 1291–1312.

Lenz, G. S., & Lawson, C. (2011). Looking the part: Television leads less informed citizens to vote based on candidates' appearance. *American Journal of Political Science*, *55*, 574–589.

Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*, 57–64.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska directed emotional faces (KDEF)*. Karolinska Institutet.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*, 1122–1135.

Ma, D. S., Correll, J., & Wittenbrink, B. (2018). The effects of category and physical features on stereotyping and evaluation. *Journal of Experimental Social Psychology*, *79*, 42–50.

Ma, D. S., Koltai, K., McManus, R. M., Bernhardt, A., Correll, J., & Wittenbrink, B. (2018). Race signaling features: Identifying markers of racial prototypicality among asians, blacks, latinos, and whites. *Social Cognition*, *36*, 603–625.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology*, *51*, 93–120.

Martinez, J. E., Funk, F., & Todorov, A. (2020). Quantifying idiosyncratic and shared contributions to stimulus evaluations. *Behavior Research Methods*, *52*, 1428–1444.

McCrae, R. R. (2002). NEO-PI-R data from 36 cultures. In *The five-factor model of personality across cultures* (pp. 105–125). Springer.

McCrae, R. R., & Costa, P. T. (2008). The five-factor theory of personality. In J. Oliver, P. Naumann, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 159–181). Guilford.

Montepare, J. M., & Zebrowitz, L. A. (1998). Person perception comes of age: The salience and significance of age in social judgments. In *Vol. 30. Advances in experimental social psychology* (pp. 93–161). Elsevier.

Montepare, J. M., & Zebrowitz McArthur, L. (1986). The influence of facial characteristics on children's age perceptions. *Journal of Experimental Child Psychology*, *42*, 303–314.

Moshontz, H., Campbell, L., Ebersole, C. R., Ijzerman, H., Urry, H. L., Forscher, P. S., et al. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*, 501–515.

Nakamura, K., & Watanabe, K. (2019). Data-driven mathematical model of East-Asian facial attractiveness: the relative contributions of shape and reflectance to attractiveness judgements. *Royal Society Open Science*, *6*, 182–189.

Oh, D., Buck, E. A., & Todorov, A. (2019). Revealing hidden gender biases in competence impressions of faces. *Psychological Science*, *30*, 65–79.

Oh, D., Dotsch, R., Porter, J., & Todorov, A. (2019). Gender biases in impressions from faces: Empirical studies and computational models. *Journal of Experimental Psychology: General*, *149*, 323–342.

Oh, D., Dotsch, R., & Todorov, A. (2019). Contributions of shape and reflectance information to social judgments from faces. *Vision Research*, *165*, 131–142.

Oh, D., Wedel, N., Labbree, B., & Todorov, A. (in preparation). Data-driven models of face-based trustworthiness judgments unconfounded by attractiveness.

Olivola, C. Y., & Todorov, A. (2010a). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, *34*, 83–110.

Olivola, C. Y., & Todorov, A. (2010b). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology*, *46*, 315–324.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*, 11087–11092.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, *49*, 197–237.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

O'Toole, A. J. (2011). Cognitive and computational approaches to face recognition. In G. Rhodes, A. Calder, M. Johnson, & J. V. Haxby (Eds.), *Oxford handbook of face perception* Oxford University Press.

O'Toole, A. J., Price, T., Vetter, T., Bartlett, J. C., & Blanz, V. (1999). 3D shape and 2D surface textures of human faces: the role of "averages" in attractiveness and age. *Image and Vision Computing*, *18*, 9–19.

Over, H., & Cook, R. (2018). Where do spontaneous first impressions of faces come from? *Cognition*, *170*, 190–200.

Paunonen, S. V. (1997). On chance and factor congruence following orthogonal Procrustes rotation. *Educational and Psychological Measurement*, *57*, 33–59.
Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd ed.). Harcourt Brace Jovanovich.
Poutvaara, P., Jordahl, H., & Berggren, N. (2009). Faces of politicians: Babyfacedness predicts inferred competence but not electoral success. *Journal of Experimental Social Psychology*, *45*, 1132–1135.
Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLoS One*, *7*, e34293.
Russell, R., Sinha, P., Biederman, I., & Nederhouser, M. (2006). Is pigmentation important for face recognition? Evidence from contrast negation. *Perception*, *35*, 749–759.
Ryali, C. K., Goffin, S., Winkielman, P., & Yu, A. J. (2020). From likely to likable: The role of statistical typicality in human social assessment of faces. *Proceedings of the National Academy of Sciences*, *117*, 29371–29380.
Ryali, C. K., & Yu, A. J. (2018). Beauty-in-averageness and its contextual modulations: A Bayesian statistical account. *Advances in Neural Information Processing Systems*.
Sadr, J., Jarudi, I., & Sinha, P. (2003). The role of eyebrows in face recognition. *Perception*, *32*, 285–293.
Said, C. P., & Todorov, A. (2011). A statistical model of facial attractiveness. *Psychological Science*, *22*, 1183–1190.
Saucier, G. (2003). Factor structure of English-language personality type-nouns. *Journal of Personality and Social Psychology*, *85*, 695–708.
Schneider, D. J. (1973). Implicit personality theory: A review. *Psychological Bulletin*, *79*, 294–309.
Scott, I. M. L., Clark, A. P., Josephson, S. C., Boyette, A. H., Cuthill, I. C., Fried, R. L., et al. (2014). Human preferences for sexually dimorphic faces may be evolutionarily novel. *Proceedings of the National Academy of Sciences*, *111*, 14388–14393.
Secord, P. F. (1958). Facial features and inference processes in interpersonal perception. In R. Tagiuri, & L. Petrullo (Eds.), *Person perception and interpersonal behavior* (pp. 300–315). Stanford University Press.
Sharma, S. (1996). *Applied multivariate techniques*. John Wiley & Sons.
Singular Inversions. (2005). *FaceGen Main Software Development Kit*. Vancouver, BC, Canada. Version 3.1.
Sofer, C., Dotsch, R., Oikawa, M., Oikawa, H., Wigboldus, D. H. J., & Todorov, A. (2017). For your local eyes only: Culture-specific face typicality influences perceptions of trustworthiness. *Perception*, *46*, 914–928.
Sofer, C., Dotsch, R., Wigboldus, D. H. J., & Todorov, A. (2015). What is typical is good: the influence of face typicality on perceived trustworthiness. *Psychological Science*, *26*, 39–47.
Sormaz, M., Young, A. W., & Andrews, T. J. (2016). Contributions of feature shapes and surface cues to the recognition of facial expressions. *Vision Research*, *127*, 1–10.
South Palomares, J. K., Sutherland, C. A. M., & Young, A. W. (2018). Facial first impressions and partner preference models: Comparable or distinct underlying structures? *British Journal of Psychology*, *109*, 538–563.
Stolier, R. M., Hehman, E., & Freeman, J. B. (2020). Trait knowledge forms a common structure across social cognition. *Nature Human Behaviour*, *4*, 361–371.
Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences*, *115*, 9210–9215.

Sussman, A. B., Petkova, K., & Todorov, A. (2013). Competence ratings in US predict presidential election outcomes in Bulgaria. *Journal of Experimental Social Psychology*, *49*, 771–775.

Sutherland, C. A. M., Burton, N. S., Wilmer, J. B., Blokland, G. A. M., Germine, L., Palermo, R., et al. (2020). Individual differences in trust evaluations are shaped mostly by environments, not genes. *Proceedings of the National Academy of Sciences*, *117*, 10218–10224.

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

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*, 105–118.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*, 25–29.

Tingley, D. (2014). Face-off: Facial features and strategic choice. *Political Psychology*, *35*, 35–55.

Todorov, A. (2009). On the richness and limitations of dimensional models of social perception. *Behavioral and Brain Sciences*, *32*, 402–403.

Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton University Press.

Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion*, *13*, 724–738.

Todorov, A., Dotsch, R., Wigboldus, D. H. J., & Said, C. P. (2011). Data-driven methods for modeling social perception: Modeling social perception. *Social and Personality Psychology Compass*, *5*, 775–791.

Todorov, A., Gobbini, M. I., Evans, K. K., & Haxby, J. V. (2007). Spontaneous retrieval of affective person knowledge in face perception. *Neuropsychologia*, *45*, 163–173.

Todorov, A., Loehr, V., & Oosterhof, N. N. (2010). The obligatory nature of holistic processing of faces in social judgments. *Perception*, *39*, 514–532.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*, 1623–1626.

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

Todorov, A., & Oosterhof, N. (2011). Modeling social perception of faces. *IEEE Signal Processing Magazine*, *28*, 117–122.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, *27*, 813–833.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*, 455–460.

Todorov, A., Uddenberg, S. D., Peterson, J. C., Griffiths, T. L., & Suchow, J. W. (2020). *Data-driven, photorealistic social face-trait encoding, prediction, and manipulation using deep neural networks*. U.S. Provisional Application for Patent No. 62903267. U.S. Patent and Trademark Office.

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology*, *83*, 1051–1065.

Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *Journal of Experimental Social Psychology*, *39*, 549–562.
Todorov, A., & Uleman, J. S. (2004). The person reference process in spontaneous trait inferences. *Journal of Personality and Social Psychology*, *87*, 482–493.
Torrance, J. S., Wincenciak, J., Hahn, A. C., DeBruine, L. M., & Jones, B. C. (2014). The relative contributions of facial shape and surface information to perceptions of attractiveness and dominance. *PLoS One*, *9*, e104415.
Toscano, H., Schubert, T. W., Dotsch, R., Falvello, V., & Todorov, A. (2016). Physical strength as a cue to dominance: A data-driven approach. *Personality and Social Psychology Bulletin*, *42*, 1603–1616.
Troje, N. F., & Bülthoff, H. H. (1996). Face recognition under varying poses: The role of texture and shape. *Vision Research*, *36*, 1761–1771.
Tucker, L. R. (1951). *A method for synthesis of factor analysis studies (Personnel Research Section Report 984)*. Educational Testing Service.
Uleman, J. S., Blader, S. L., & Todorov, A. (2005). Implicit impressions. In R. Hassin, J. S. Uleman, & J. A. Bargh (Eds.), *The new unconscious* (pp. 362–392).
Verosky, S. C., Porter, J. M., Martinez, J. E., & Todorov, A. (2018). Robust effects of affective person learning on evaluation of faces. *Journal of Personality and Social Psychology*, *114*, 516–528.
Verosky, S. C., & Todorov, A. (2010). Generalization of affective learning about faces to perceptually similar faces. *Psychological Science*, *21*, 779–785.
Verosky, S. C., & Todorov, A. (2013). When physical similarity matters: Mechanisms underlying affective learning generalization to the evaluation of novel faces. *Journal of Experimental Social Psychology*, *49*, 661–669.
Verosky, S. C., Todorov, A., & Turk-Browne, N. B. (2013). Representations of individuals in ventral temporal cortex defined by faces and biographies. *Neuropsychologia*, *51*, 2100–2108.
Walker, M., Jiang, F., Vetter, T., & Sczesny, S. (2011). Universals and cultural differences in forming personality trait judgments from faces. *Social Psychological and Personality Science*, *2*, 609–617.
Walker, M., & Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision*, *9*, 1–13.
Walker, M., & Vetter, T. (2016). Changing the personality of a face: Perceived Big Two and Big Five personality factors modeled in real photographs. *Journal of Personality and Social Psychology*, *110*, 609–624.
Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology*, *37*, 395–412.
Wiggins, J. S., Phillips, N., & Trapnell, P. (1989). Circular reasoning about interpersonal behavior: Evidence concerning some untested assumptions underlying diagnostic classification. *Journal of Personality and Social Psychology*, *56*, 296–305.
Wiggins, J. S., & Pincus, A. L. (1992). Personality: Structure and assessment. *Annual Review of Psychology*, *43*, 473–504.
Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*, 592–598.
Wojciszke, B. (1994). Multiple meanings of behavior: Construing actions in terms of competence or morality. *Journal of Personality and Social Psychology*, *67*, 222–232.
Xie, S. Y., Flake, J. K., & Hehman, E. (2018). Perceiver and target characteristics contribute to impression formation differently across race and gender. *Journal of Personality and Social Psychology*, *117*, 364–385.
Young, A. W. (2018). Faces, people and the brain: the 45th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology*, *71*, 569–594.

Zebrowitz, L. A. (1997). *Reading faces: Window to the soul? New directions in social psychology*. Routledge.
Zebrowitz, L. A. (2017). First impressions from faces. *Current Directions in Psychological Science*, *26*, 237–242.
Zebrowitz, L. A., & Collins, M. A. (1997). Accurate social perception at zero acquaintance: The affordances of a Gibsonian approach. *Personality and Social Psychology Review*, *1*, 204–223.
Zebrowitz McArthur, L., & Apatow, K. (1984). Impressions of baby-faced adults. *Social Cognition*, *2*, 315–342.
Zhang, L., Holzleitner, I. J., Lee, A. J., Wang, H., Han, C., Fasolt, V., et al. (2019). A data-driven test for cross-cultural differences in face preferences. *Perception*, *48*, 487–499.