Understanding how people make trait attributions from faces

Thesis by

Chujun Lin

In Partial Fulfillment of the Requirements for

the degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2019

(Defended May 17, 2019)

© 2019

Chujun Lin

ORCID: 0000-0002-7605-6508

ACKNOWLEDGEMENTS

Firstly, I would like to thank my beloved parents, for supporting me to pursue a PhD here at Caltech seven thousand miles away from home, for always reminding me to care about what is happening around the world and to think about the society and people around me instead of only focusing on my own little world, and for understanding my decision to pursue a career and life that is very different from what the social norm of our culture expects me to do, but that is what I am truly passionate about.

I would like to express my sincere gratitude to my two extraordinary advisors, Ralph Adolphs and R. Michael Alvarez, who are always excited to listen to me to talk about my ideas, who always provide me with very helpful advice on my research, who teach me to be a responsible scientist and do transparent and high-quality research, who know my personality well and tailor their mentoring to my personality, and who not only care about my work, but also care about whether I am happy and healthy in life.

I am grateful to my amazing friends, with a special mention to Tao Liang, Bow Nicha Leethochawalit, Ashish Goel, Tung Pakorn Wongwaitayakornkul, and Aaron Mendelovitz, who have always been there through my ups and downs during these years at Caltech, with whom I have shared so many unforgettable adventures in the mountains and the seas, and who know my weakness and weirdness and accept me for who I am. I am also grateful to my friends from the Caltech Theater Arts and Caltech Badminton Club, for all the fun and hard work we have shared.

My sincere thanks also go to Prof. Colin Camerer, Prof. Dean Mobbs, and Prof. Antonio Rangel, from whom I have received insightful advice on my research and my career plan,

and from whom I have learnt new perspectives on designing interesting studies and improving work efficiency.

I thank my first-year classmates in the social science program, with a special mention to George Vega Yon, Kevin Laughren, and Alejandro Robinson-Cortés, for the stressful days we were working together in preparation for the preliminary examination.

I thank my fellow labmates in the Adolphs Lab, with a special mention to Umit Keles, Remya Nair, Anita Tusche, and Shuo Wang. It was fantastic to have had the opportunity to work with you all.

Last but not least, I would like to thank the staff at the Division of Humanities and Social Sciences, with a special mention to Laurel M. Auchampaugh and Christopher J. Birtja, who have always been super patient and helpful with all my questions.

People always say grad school is hard. I did feel the hard work; but because of all of you, my grad-school life at Caltech has been filled with joy and excitement. Thank you so much!

ABSTRACT

This thesis is motivated by the fascinating question of how people make inferences about others from their faces. How do we infer somebody's intent or personality merely from looking at them? I studied this question by investigating how people make trait attributions in two specific contexts—political election (Chapter 2) and political corruption (Chapter 3)—as well as how people make a large variety of trait attributions from faces in general (Chapter 4). I employed novel methods to representatively sample the words used to rate faces, and to select the facial stimuli themselves (e.g., using artificial neural networks), to test the reproducibility and generalizability of my results (e.g., pre-registration, generalization across participants from different cultures), and to elucidate the underlying mechanisms (e.g., mediation modeling, digital manipulation of facial stimuli). The results demonstrated that trait attributions from politician's faces were associated with real election outcomes in different cultures, and that culture shaped trait attributions relevant to a given context (Chapter 2); trait attributions from politician's faces were also associated with real corruption/violation records of the politicians, and perceived corruptibility was associated with the width of the face (Chapter 3). Trait attributions from faces in general (Chapter 4) were well-described by four novel dimensions that I discovered: critical/condescending, leadership/competence, female-stereotype, and youth-stereotype. Taken together, the findings provide a new psychological framework for trait attributions, demonstrate cross-cultural generalizability, and link trait attributions to real-world behaviors.

PUBLISHED CONTENT AND CONTRIBUTIONS

Lin, C., Adolphs, R., & Alvarez, R. M. (2017). Cultural effects on the association between election outcomes and face-based trait inferences. *PloS one*, 12(7). doi: 10.1371/journal.pone.0180837.

C.L contributed to developing the concept of the project, designing and programming the experiment, collecting the data, analyzing the data, and drafting and revising the manuscript.

Lin, C., Adolphs, R., & Alvarez, R. M. (2018). Inferring whether officials are corruptible from looking at their faces. *Psychological science*, 29(11), 1807-1823. doi: 10.1177/0956797618788882.

C.L contributed to developing the concept of the project, designing and programming the experiment, collecting the data, analyzing the data, and drafting and revising the manuscript.

Lin, C., Keles, U., & Adolphs, Ralph. (2019). The comprehensive space for trait attributions from faces is four dimensional. [Under review at *Nature Human Behavior*].

C.L contributed to developing the concept of the project, designing and programming the experiment, collecting the data, analyzing the data, and drafting and revising the manuscript.

TABLE OF CONTENTS

Chapter 1

Literature Review

1.1 Why is trait attribution from faces important to study?

Faces convey important social information about their owners, such as gender, age, race, identity, and emotional states. Intriguingly, people infer numerous other traits from viewing a face, including relatively stable traits regarding a person's intent (e.g., threatening, trustworthy), ability (e.g., competent, intelligent), and personality (e.g., outgoing, aggressive)[1,2]. While many, if not most, of these trait inferences are inaccurate[3], some recent work has demonstrated that certain trait attributions are associated with real-world social outcomes. For example, political candidates whose faces looked more competent to participants in laboratory experiments (who knew nothing else about the politicians) were more likely to be the actual winners in past elections[4,5]; business managers whose faces looked more powerful were more likely to be the leaders of companies that earned higher profits[6,7]; and innocent individuals whose faces looked less trustworthy were more likely to be assigned harsher sentences before they were exonerated from the charges[8]. These findings suggest that, regardless of whether trait attributions from faces are accurate or not, they might shape important social behaviors, which has motivated a large literature to understand how people make trait attributions from faces and what consequences they have.

1.2 What is special about trait attributions from faces?

First, people make trait attributions from faces rapidly and automatically. Behavioral studies asking participants to make trait attributions after viewing faces for a range of brief exposure

times (17, 26, 33, 39, 50, 67, 100, 167, 500, 1000, 1700 milliseconds) showed that attributions of threat made after the faces were exposed for 39ms were highly correlated with those made after longer exposure times (r = 0.77)[9], and that attributions of trustworthiness made after the faces exposed for 33ms and longer were significantly correlated with those made under unlimited exposure times (*rs* increased from .22 to .79)[10,11]. Neuroimaging studies using functional magnetic resonance imaging to record neural activations in the amygdala while participants were viewing faces revealed increased responses in both left and right amygdala to less trustworthy-looking faces even when participants were not making attributions of trustworthiness explicitly (participants were told they were performing a memory task)[12].

Second, these rapid and automatic trait attributions from faces powerfully shape behavior even when more reliable information than the face is present. For example, in trust games where participants (both children and adults) decided the amount of money to invest in each trustee, trustworthy-looking trustees received substantially greater investments than untrustworthy-looking trustees. This effect was observed both when the trustee's face was the only information available[13] and when the trustee's face as well as reputational information were available to the participants[14].

Third, individuals across different cultures and of different ages make highly similar trait attributions from faces. In cross-cultural studies where participants were asked to freely describe faces, the most frequently used words were found to be similar between participants from east-Asian and Western cultures; similar ratings were also found when participants from different cultures were asked to make attributions of given traits based on faces[1,15]. Developmental studies comparing attributions of trustworthiness, dominance, and

competence made by children and adults demonstrated that children as young as age 3 made attributions that converged with those of adults[16,17].

While trait attributions based on the same facial image are highly consensual across individuals, trait attributions of the same facial identity based on his/her different images or under different contexts can be highly variable. Previous studies using different facial images of the same facial identity that varied in facial expressions (any spontaneous facial expression) found that trait ratings (averaged across participants) of the same facial identity can vary substantially from image to image (e.g., correlations of creativity attributions between different images of the same facial identity ranged from .13 to .67 across facial identities; the low correlation was not due to low reliability of the attribution), and that different images of the same facial identity were favored under different contexts (e.g., consultant applications, online dating, Facebook photos, political campaigns, and film auditions)[3,18]. Building on these findings about contextual differences of trait attributions, Chapter 2 and Chapter 3 of this thesis focus on two interesting contexts—a positive context (competence) and a negative context (corruption)—and investigate how people make trait attributions of politicians under these two specific contexts.

Finally, the number of trait attributions that people can make from faces is large. Previous studies asking participants to freely describe facial identities based on photos with a neutral expression showed that participants (sample sizes ranging from 20 to 55) were able to provide more than 1,100 different words and phrases to describe the faces (stimulus-set size ranging from 60 to 66), which spanned attributions of demographic characteristics (e.g., gender, age), physical appearance (e.g., pretty, healthy), social evaluation (e.g., trustworthy, competent), personality (e.g., quiet, sociable), emotion (e.g., happy, sad), and preferences

(e.g., "hates sports")[1,2]. Compared to the large number of trait attributions people are able to make from faces, the number of traits examined in prior laboratory studies is much smaller (typically 13 to 25 traits)[19,20], and modern psychological models of trait attributions from faces are based on the inspection of merely 13 traits[2,21]. Aiming to break through this limitation of previous research, Chapter 4 of this thesis integrates a an inclusive set of traits and uses a novel data-driven approach to characterize the dimensionality of trait attributions from faces.

1.3 How do people make trait attributions from faces?—what we know and don't know.

How do people make inferences about a person's intent, ability, and personality from merely the face? Presumably, the underlying mechanism will vary depending on the specific traits (e.g., traits related to physical appearance such as feminine versus personality traits such as conscientious; traits belonging to the intent dimension versus traits belonging to the ability dimension) and the specific context of the attributions (e.g., making attributions under time pressure versus having more time to look at the face; making attributions from faces of different genders or different races; making attributions for the purpose of electing a politician versus choosing a romantic partner). While, as mentioned above, there are still a large number of traits and a wide range of contexts that remain to be investigated, prior research has uncovered three (non-mutually exclusive) mechanisms so far, which I detail in the following sections.

1.3.1 Specific facial features facilitate trait attribution from faces

First, prior studies have shown that trait attributions from emotionally neutral faces rely on visual cues of facial structure, skin texture, and skin tone. A key technique for identifying the visual cues that facilitate the attribution of a trait is using reverse correlation[22]. This technique asks participants to choose the face that best exhibits a trait between pairs of faces that were generated with the same base face (e.g., a face averaged over a large number of grey-scale faces of the same gender and race) but superimposed with complementary noise patterns (a random noise pattern versus its inverted noise pattern). By averaging the selected noise patterns across multiple trials, the reverse correlation technique has revealed that brighter mouth and eyebrow regions, and darker eye and hair regions, facilitate attributions of trustworthiness, and that brighter hair and cheekbone regions and darker eyebrow and chin regions facilitate attributions of dominance[22]. While reverse correlation is informative for identifying the regions and shades of a face that contribute to trait attributions, it is still unclear which specific facial features are the causal determinant of this contribution (e.g., both the structural feature—a prominent eyebrow bridge, and the facial hair feature—thick eyebrows, could make the eyebrow region look darker in a grey-scale image). Importantly, reverse correlation presents viewers with highly artificial images, and so it is not clear that they use the same features when looking at real faces.

Another approach, with perhaps more external validity, is to manipulate specific facial features to test their causal effect on trait attribution (e.g., facial width-to-height ratio, eye size, skin texture and color, facial symmetry and averageness). For example, by digitally manipulating the width-to-height ratio of the same facial stimuli, prior research showed that men with higher facial width-to-height ratio were perceived as less trustworthy[23]. Furthermore, by warping individual faces to a same-sex averaged face, researchers were able

to digitally manipulate the averageness of the individual faces (i.e., the difference between an individual face and the averaged face) or to isolate texture-difference from shape-difference of the individual faces, which revealed that faces with greater averageness and more homogeneous texture were perceived as healthier and more attractive[24–26].

However, the fundamental challenge to this approach for understanding how people make trait attributions from specific facial features is that processing of faces (in particular when forming rapid, unreflective first impressions) might not engage decomposition of facial features but rather require holistic processing of the faces as wholes (engaging interactions, perhaps nonlinear, between multiple features at once). This holistic processing of faces is borne out by a large body of research on facial identity recognition[27–29]. In the case of trait attribution from faces, even if we observe that manipulations of specific facial features causally change the attribution of a given trait (e.g., darkening the eyebrow region increases perceived dominance, or increasing the facial width-to-height ratio decreases perceived trustworthiness), it is not straightforward to conclude that those specific facial features themselves are part of the underlying mechanism of trait attribution. It is possible that the manipulated changes in specific facial features might not be processed individually, and instead that it is the unintended resulting changes in the overall impressions of the face (i.e., holistic features, such as femininity of the face) that are perceived, which might presumably be the real underlying causal factor of trait attribution. Changing any part of a face influences how we see other parts of the face—and it is these contextual effects on other parts that could be the main mechanism driving the trait attribution.

1.3.2 Trait attribution from faces is an overgeneralization

A second theory that explains how people make trait attributions from faces, related to the challenge of holistic processing, is that trait attributions from emotionally neutral faces may be overgeneralizations of responses to groups (babies, unfamiliar others) and cues (emotional expressions) that may have been important to recognize quickly in evolution[30]. For example, the baby-face overgeneralization hypothesis posits that the benefit of rapidly identifying babies—who are intellectually and physically weak and need others' care—might have reinforced our response to facial features of babies; furthermore, the comparatively low cost of mistakenly caring for a baby-faced individual (who is not a baby) versus failing to care for a baby, might have produced a strong tendency to attribute baby-like traits to baby-faced individuals over evolution[31–33]. This hypothesis is supported by the evidence that artificial neural networks trained to discriminate faces of babies from those of adults predicted human subject's attributions of baby-like traits (e.g., warm, not strong) to baby-faced adults[32].

Similarly, responding appropriately to emotional expressions has important adaptive value (e.g., avoiding individuals with facial expressions of anger). Prior studies analyzing the correlations between trait attributions of emotionally neutral faces (based on judgments made by human subjects) and the probability that the faces got classified into one of six basic emotion categories (based on a computer vision algorithm built for detecting emotional expressions) found that, for example, faces that were perceived to exhibit positive traits (e.g., sociable, responsible, caring) were more likely to be classified as expressing happiness, faces that were perceived to exhibit negative traits (e.g., mean, unhappy) were more likely to be classified as expressing anger and disgust, and faces that were perceived to be more dominant

were more likely to be classified as expressing anger while less likely to be classified as expressing surprise or sad[2,34,35].

While the evolutionary pressure on overgeneralizing responses to critical groups and facial cues might explain why trait attributions from faces are spontaneous and show high consensus across individuals, the overgeneralization mechanism alone might not be sufficient to explain how people make trait attributions from faces. For example, the baby-facedness overgeneralization theory predicts that baby-faced individuals will be perceived as weak and incompetent, and these negative trait impressions might prevent their success as business leaders (or one could hypothesize the opposite, because they will be perceived as warm and trustworthy). However, prior research suggests that the effect of baby-facedness on success varies by race (positive for black leaders but negative for white leaders)[36]. This finding suggests that learned stereotypes (racial stereotypes) might integrate with innate overgeneralizations in determining trait attributions from faces.

1.3.3 Trait attribution from faces engages both bottom-up and top-down processes

Third, related to the debate between innate overgeneralization (from facial features) and learned stereotypes, prior studies have shown that trait attributions from emotionally neutral faces not only engage perception of facial features ("bottom-up" processes; e.g., facial structure) but are also shaped by an understanding of the trait being judged ("top-down" processes; e.g., learned stereotypes related to the trait)[37–43]. For example, facial features of Blacks will not only activate learned stereotypes associated with them (e.g., low socioeconomic status), but also facilitate stereotype-congruent predictions which in turn bias the perception of facial features as well as the attribution of traits[37]. For example, perception of facial lightness from racially-ambiguous faces was found to be biased by the race label

the face was given (lighter for faces labeled white and darker for faces labeled black)[44]; and attribution of race to racially-ambiguous faces was found to be biased by the clothes the person was wearing (high-status attire facilitated attributions of White while low-status attire facilitated attributions of Black)[45]. However, while the argument that "top-down" processes bias trait attributions (e.g., judgement of race from faces) is widely agreed upon in the literature, it is still under debate whether "top-down" processes bias perception *per se* (e.g., perception of facial features)[46].

Besides learned stereotypes, people's understanding of the traits being judged might is of course also shaped by semantic knowledge (yet another "top-down" process). For example, in the context of selecting an employee, the attribution of trustworthiness might be closely related to how responsible the individual is; whereas, in the context of selecting a romantic partner, the attribution of trustworthiness might be more about how honest the individual is. Therefore, in one context the attribution of trustworthiness from faces might show higher correlation with the attribution of responsibility, while in another context it might show higher correlation with honesty. In fact, some studies[47–49] have questioned whether responses elicited by lexical stimuli (e.g., traits describing people, words describing emotions) for a psychological domain (e.g., person perception from faces, emotion categorization from texts) reflect the structure of the psychological domain (e.g., face-evaluation dimensions, emotion dimensions) or the structure of the lexical stimuli (e.g., semantic similarity between the traits, scenario similarity between the texts). I address this important question in one of the analyses presented in Chapter 4.

1.4 Thesis Overview

Building on previous literature, this thesis provides novel studies into how people make inferences about people's traits from their faces in specific contexts (Chapter 2 and Chapter 3) and in general (Chapter 4). While previous research showed that trait attributions from faces are associated with election outcomes[4,20], little was known about how culture might influence trait attributions from faces of political candidates. Chapter 2 examines how Caucasian participants and Korean participants make trait attributions from faces of political candidates, and how those trait attributions are associated with real election outcomes in the United States and South Korea. While previous research on face evaluation in the context of politics has focused on positive outcomes (e.g., electoral success), trait attributions from faces might also be consequential for negative outcomes (e.g., political corruption). Chapter 3 compares trait attributions from faces of politicians who had records of corruption/violation with those who had clean records, and discovers a particular facial features that partly accounts for why some politicians look more corruptible than others.

When given a context (e.g., deciding which political candidate to vote for, deciding whether an individual who has been charged with a crime is guilty or innocent), people might make attributions of only a small number of traits from faces that are relevant to the context. However, when there is no context and people are free to form any impression from faces, the number of trait attributions people make from faces is large in general[1,2]. Chapter 4 employs novel methods (natural language processing and state-of-the-art computer vision processing) to derive an inclusive and non-redundant set of traits and faces, which enables a comprehensive characterization of the dimensionality of trait attributions from faces and its generalizability across countries. This provided discovery of a novel dimensional space that I propose characterizes person perception from faces.

1.5 References

1. Sutherland, C. A. M. *et al.* Facial First Impressions Across Culture: Data-Driven Modeling of Chinese and British Perceivers' Unconstrained Facial Impressions. *Pers. Soc. Psychol. Bull.* **44**, 521–537 (2018).
2. Oosterhof, N. N. & Todorov, A. The functional basis of face evaluation. *Proc. Natl. Acad. Sci.* **105**, 11087–11092 (2008).
3. Todorov, A. *Face value: The irresistible influence of first impressions*. (Princeton University Press, 2017).
4. Todorov, A. Inferences of Competence from Faces Predict Election Outcomes. *Science* **308**, 1623–1626 (2005).
5. Ballew, C. C. & Todorov, A. Predicting political elections from rapid and unreflective face judgments. *Proc. Natl. Acad. Sci.* **104**, 17948–17953 (2007).
6. Rule, N. O. & Ambady, N. The Face of Success: Inferences from Chief Executive Officers' Appearance Predict Company Profits. *Psychol. Sci.* **19**, 109–111 (2008).
7. Rule, N. O. & Ambady, N. Face and fortune: Inferences of personality from Managing Partners' faces predict their law firms' financial success. *Leadersh. Q.* **22**, 690–696 (2011).

8. Wilson, J. P. & Rule, N. O. Facial Trustworthiness Predicts Extreme Criminal-Sentencing Outcomes. *Psychol. Sci.* **26**, 1325–1331 (2015).

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

10. Todorov, A., Pakrashi, M. & Oosterhof, N. N. Evaluating Faces on Trustworthiness After Minimal Time Exposure. *Soc. Cogn.* **27**, 813–833 (2009).

11. Willis, J. & Todorov, A. First impressions: making up your mind after a 100-ms exposure to a face. *Psychol. Sci.* **17**, 592–598 (2006).

12. Engell, A. D., Haxby, J. V. & Todorov, A. Implicit Trustworthiness Decisions: Automatic Coding of Face Properties in the Human Amygdala. *J. Cogn. Neurosci.* **19**, 1508–1519 (2007).

13. Ewing, L., Caulfield, F., Read, A. & Rhodes, G. Perceived trustworthiness of faces drives trust behaviour in children. *Dev. Sci.* **18**, 327–334 (2015).

14. Rezlescu, C., Duchaine, B., Olivola, C. Y. & Chater, N. Unfakeable Facial Configurations Affect Strategic Choices in Trust Games with or without Information about Past Behavior. *PLOS ONE* **7**, e34293 (2012).

15. Rule, N. O. *et al.* Polling the face: Prediction and consensus across cultures. *J. Pers. Soc. Psychol.* **98**, 1–15 (2010).

16. Cogsdill, E. J., Todorov, A. T., Spelke, E. S. & Banaji, M. R. Inferring Character From Faces: A Developmental Study. *Psychol. Sci.* **25**, 1132–1139 (2014).

17. Antonakis, J. & Dalgas, O. Predicting Elections: Child's Play! *Science* **323**, 1183–1183 (2009).

18. Todorov, A. & Porter, J. M. Misleading First Impressions: Different for Different Facial Images of the Same Person. *Psychol. Sci.* **25**, 1404–1417 (2014).

19. Hehman, E., Sutherland, C. A. M., Flake, J. K. & Slepian, M. L. The unique contributions of perceiver and target characteristics in person perception. *J. Pers. Soc. Psychol.* **113**, 513–529 (2017).

20. Olivola, C. Y. & Todorov, A. Elected in 100 milliseconds: Appearance-Based Trait Inferences and Voting. *J. Nonverbal Behav.* **34**, 83–110 (2010).

21. Sutherland, C. A. M. *et al.* Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition* **127**, 105–118 (2013).

22. Dotsch, R. & Todorov, A. Reverse Correlating Social Face Perception. *Soc. Psychol. Personal. Sci.* **3**, 562–571 (2012).

23. Stirrat, M. & Perrett, D. I. Valid Facial Cues to Cooperation and Trust: Male Facial Width and Trustworthiness. *Psychol. Sci.* **21**, 349–354 (2010).

24. Grammer, K. & Thornhill, R. Human (Homo sapiens) facial attractiveness and sexual selection: The role of symmetry and averageness. *J. Comp. Psychol.* **108**, 233–242 (1994).

25. Fink, B., Grammer, K. & Thornhill, R. Human (Homo sapiens) facial attractiveness in relation to skin texture and color. *J. Comp. Psychol.* **115**, 92–99 (2001).

26. Rhodes, G. *et al.* Do facial averageness and symmetry signal health? *Evol. Hum. Behav.* **22**, 31–46 (2001).

27. Peterson, M. A. & Rhodes, G. *Perception of Faces, Objects, and Scenes: Analytic and Holistic Processes*. (Oxford University Press, 2006).

28. Young, A. W., Hellawell, D. & Hay, D. C. Configurational Information in Face Perception. *Perception* **16**, 747–759 (1987).

29. James W. Tanaka & Martha J. Farah. Parts and Wholes in Face Recognition. *Q. J. Exp. Psychol.* **46A**, 225–245 (1993).

30. Zebrowitz, L. A. First Impressions From Faces. *Curr. Dir. Psychol. Sci.* **26**, 237–242 (2017).

31. Zebrowitz, L. A. & Montepare, J. M. Social Psychological Face Perception: Why Appearance Matters. *Soc. Personal. Psychol. Compass* **2**, 1497 (2008).

32. Zebrowitz, L. A., Fellous, J., Mignault, A. & Andreoletti, C. Trait Impressions as Overgeneralized Responses to Adaptively Significant Facial Qualities: Evidence from Connectionist Modeling. *Personal. Soc. Psychol. Rev.* **7**, 194–215 (2003).

33. Zebrowitz, L. A., Bronstad, P. M. & Lee, H. K. THE CONTRIBUTION OF FACE FAMILIARITY TO INGROUP FAVORITISM AND STEREOTYPING. in (2007).

34. Said, C. P., Sebe, N. & Todorov, A. *BRIEF REPORTS Structural Resemblance to Emotional Expressions Predicts Evaluation of Emotionally Neutral Faces*. (2009).

35. Montepare, J. M. & Dobish, H. The Contribution of Emotion Perceptions and Their Overgeneralizations to Trait Impressions. *J. NONVERBAL Behav.* 18

36. Livingston, R. W. & Pearce, N. A. The Teddy-Bear Effect: Does Having a Baby Face Benefit Black Chief Executive Officers? *Psychol. Sci.* **20**, 1229–1236 (2009).

37. Brooks, J. A. & Freeman, J. B. Neuroimaging of person perception: A social-visual interface. *Neurosci. Lett.* **693**, 40–43 (2019).

38. Stolier, R. M., Hehman, E. & Freeman, J. B. A Dynamic Structure of Social Trait Space. *Trends Cogn. Sci.* **22**, 197–200 (2018).

39. Stolier, R. M. & Freeman, J. B. Functional and Temporal Considerations for Top-Down Influences in Social Perception. *Psychol. Inq.* **27**, 352–357 (2016).

40. Freeman, J. B., Stolier, R. M., Brooks, J. A. & Stillerman, B. S. The neural representational geometry of social perception. *Curr. Opin. Psychol.* **24**, 83–91 (2018).

41. Stolier, R. M. & Freeman, J. B. A Neural Mechanism of Social Categorization. *J. Neurosci.* **37**, 5711–5721 (2017).

42. Mason, M. F., Cloutier, J. & Macrae, C. N. On Construing Others: Category and Stereotype Activation from Facial Cues. *Soc. Cogn.* **24**, 540–562 (2006).

43. Freeman, J. B. & Johnson, K. L. More Than Meets the Eye: Split-Second Social Perception. *Trends Cogn. Sci.* **20**, 362–374 (2016).

44. Levin, D. T. & Banaji, M. R. Distortions in the perceived lightness of faces: The role of race categories. *J. Exp. Psychol. Gen.* **135**, 501–512 (2006).

45. Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M. & Ambady, N. Looking the Part: Social Status Cues Shape Race Perception. *PLOS ONE* **6**, e25107 (2011).

46. Firestone, C. & Scholl, B. J. Cognition does not affect perception: Evaluating the evidence for 'top-down' effects. *Behav. Brain Sci.* **39**, e229 (2016).

47. Kuusinen, J. Factorial invariance of personality ratings. *Scand. J. Psychol.* **10**, 33–44 (1969).

48. Mulaik, S. A. Are personality factors raters' comceptual factors? *J. Consult. Psychol.* **28**, 506–511 (1964).

49. Bruner, J. S. & Tagiuri, R. The perception of people. in *Handbook of social psychology* **2**, (Addison Wesley, 1954).

Chapter 2

Cultural effects on the association between election outcomes and face-based trait inferences

How competent a politician looks, as assessed in the laboratory, is correlated with whether the politician wins in real elections. This finding has led many to investigate whether the association between candidate appearances and election outcomes transcends cultures. However, these studies have largely focused on European countries and Caucasian candidates. To the best of our knowledge, there are only four cross-cultural studies that have directly investigated how face-based trait inferences correlate with election outcomes across Caucasian and Asian cultures. These prior studies have provided some initial evidence regarding cultural differences, but methodological problems and inconsistent findings have complicated our understanding of how culture mediates the effects of candidate appearances on election outcomes. Additionally, these four past studies have focused on positive traits, with a relative neglect of negative traits, resulting in an incomplete picture of how culture may impact a broader range of trait inferences.

To study Caucasian-Asian cultural effects with a more balanced experimental design, and to explore a more complete profile of traits, here we compared how Caucasian and Korean participants' inferences of positive and negative traits correlated with U.S. and Korean election outcomes. Contrary to previous reports, we found that inferences of competence (made by participants from both cultures) correlated with both U.S. and Korean election outcomes. Inferences of open-mindedness and threat, two traits neglected in previous cross-cultural studies, were correlated with Korean but not U.S. election outcomes. This differential effect was found in trait judgments made by both Caucasian and Korean participants. Interestingly, the faster the participants made face-based trait inferences, the more strongly those inferences were correlated with real election outcomes. These findings provide new insights into cultural effects and the difficult question of causality underlying the association between facial inferences and election outcomes. We also discuss the implications for political science and cognitive psychology.

2.1 Introduction

Numerous studies have reported that trait inferences made by participants who had no previous knowledge of the political candidates, and who looked at the candidates' photos for as briefly as 100 milliseconds, correlate with real election outcomes [1-4]. This was initially studied for Australian elections [5], and was made most popular by the later studies for U.S. elections [1-3]. Subsequent research has examined how these face-based trait evaluations might be associated with election outcomes in other countries than the U.S. Supportive evidence reinforcing the original results has been found in Britain [6-9], Germany [10-11], France [12], Finland [13] (but see [14]), Ireland [15], Switzerland [16], Bulgaria [17], Denmark [18], Italy [19], Australia [9, 20], New Zealand [9], Brazil [21], Mexico [21], Japan [22], China [23], and Taiwan (ROC) [24, 25] (but see [26] for insignificant effects found in South Korea). Facial inferences made by both human subjects (cf. citations above) and computer algorithms [27] have been demonstrated to associate with election outcomes across cultures. Additional studies have extended this literature by exploring the association between election outcomes and a broader range of facial attributions, such as smile intensity in photos [28], and facial cues that reveal candidates' political affiliations [29, 30] and personality (e.g., extraverted/enthusiastic and disorganized/careless) [3]. Our present study focused on direct comparisons between Caucasian and Asian cultures, and on traits that are closely related to the initial study [5] (e.g., competence).

Understanding cultural effects advances our knowledge about how candidate appearances associate with election outcomes, which could have complex explanations. For instance, one study [28] has found that American politicians show more excited smiles in

their official photos than do Chinese/Taiwanese politicians. Cultural nuances such as this might mediate how face-based trait inferences correlate with real election outcomes because the inferences of traits are influenced by the perceptions of emotional expressions of the faces [31-35]. Differences in what traits are valued across cultures [36, 37] might be yet another contributing factor.

Although there are numerous studies of this topic across a range of cultures (cf. citations above), few have directly compared Caucasian and Asian cultures in the same study. To the best of our knowledge, there are only four cross-cultural studies that have directly investigated how face-based trait inferences made by human subjects correlate with real election outcomes, and made explicit comparisons between Caucasian and Asian cultures [22, 24-26]. One of these studies [22] found that inferences of power traits (dominance and facial maturity) correlated with U.S. but not Japanese election outcomes, while inferences of warmth traits (likeability and trustworthiness) correlated with Japanese but not U.S. election outcomes. Such different trait-election associations were observed for inferences made by both Caucasian and Asian participants. However, while the stimuli for U.S. candidates used in this study were winners and runner-ups in matched electoral races, the stimuli for Japanese candidates were not matched (winners and losers were from different electoral races). Another of these studies [26] found that inferences of competence correlated much more strongly with U.S. election outcomes than with Korean election outcomes. The candidates who were perceived as more competent by their Korean participants won in 61.92% of the U.S. elections but in only 49.98% of the Korean elections (which was below chance). The candidates who were perceived as more competent by their U.S. participants won in 60.31% and 52.85% of the electoral races in the U.S. and Korea, respectively. However, this study

[26] counterbalanced the ordering of image groups only for Caucasian participants but not for Korean participants; thus, all Korean participants evaluated U.S. candidates first, introducing possibly confounding order effects into the study.

Caucasian-Asian cultural differences were also studied by [24, 25], which compared U.S. and Taiwan elections. Unlike [22], [24] found that inferences of trustworthiness (one of the two domains of warmth traits) made by Caucasian participants were correlated with neither U.S. nor Taiwan election outcomes, while those made by Asian participants were negatively correlated with U.S. election outcomes. Counter to the argument in [26] that trait inferences were less important in Asian cultures and should be less associated with Asian elections, [24, 25] showed that inferences of some traits (e.g., social competence) correlated even more strongly with Taiwan than U.S. election outcomes. While [26] was a within-subject design, [22, 24-25] were not: participants in the latter two studies evaluated only ten pairs of faces from each culture, randomly chosen from the image pool. The discrepancies in the findings among these four studies, and the unbalanced experimental designs they used, complicate our understanding of how Caucasian and Asian cultural effects might mediate the association between appearance-based trait inferences and real election outcomes. To help clarify this issue was one motivation of our present study.

In addition to the lack of consensus on cultural effects (of both participants and election locations) as reviewed above, there is a second aspect of this topic that remains under-investigated: negative facial cues. While a few studies with Caucasian politicians have found that negative traits inferred from faces are strongly associated with election outcomes [3, 38-40], more attention has been given to investigating positive traits such as warmth, competence, trustworthiness and dominance [41-44], which tend to be strongly inter-

correlated. All the four Caucasian-Asian cross-cultural studies above [22, 24-26] examined only positive traits. This gap in the study of how negative traits might influence voter decisions is important because positive and negative traits could influence voter decisions through distinct mechanisms [38]. Negative advertising has been employed in political campaigns for decades. Lyndon Johnson's landslide victory in the 1964 U.S. presidential election is believed to owe much to the "Daisy" advertisement that attacked his opponent Barry Goldwater for being militarily aggressive. In the 2016 presidential election campaigns, Donald Trump questioned the intelligence of his rival Jeb Bush, and Republican attack-advertisements portrayed Hillary Clinton as a liar. Such anecdotal evidence, together with findings in [3, 38-40] suggest that it is important to understand how inferences of a variety of negative traits influence voting, and how culture mediates these effects. A second motivation for our present study was thus to provide a more comprehensive investigation of both positive and negative traits in a cross-cultural context.

To provide a more balanced experimental design (in participants, stimuli, and procedures) for studying cross-cultural effects, and to investigate multiple positive and negative traits, we asked Caucasian and Korean participants to make inferences of competence, open-mindedness, threat, and corruption for pairs of real political candidates from past U.S. and Korean elections. We found that the traits that were most strongly associated with election outcomes differed between U.S. and Korean elections, but that the associations were consistent across both Caucasian and Korean participants. These results provide new insights into the difficult question of causality underlying the association between face-based trait inferences and election outcomes. They also have implications for

studies of political behavior: they suggest that it is important to include both candidate traits and their cultural backgrounds in the classic vote choice model.

2.2 Materials and Methods

Participants

Caucasian participants (N = 40; 20 male; Age (M = 31, SD = 6.9)) and Korean participants (N = 40; 20 male; Age (M = 29, SD = 6.4)) with normal (or corrected-to-normal) vision were recruited from the general population in Southern California in early 2016. All Caucasian participants self-reported as "White, non-Hispanic" in the prescreening survey. All Korean participants were recruited through Korean-language advertisements. To balance the two subject pools, we recruited both Caucasians and Koreans from nearby colleges, churches, and through similar websites (e.g., Craigslist and Reddit for Caucasians, and Radiokorea for Koreans) (S1 Table). Based on earlier work by [1] (SOM), we established that a sample size of forty participants from each cultural background would be necessary; their study showed that the average individual accuracy of face-based competence inferences predicting U.S. election outcomes increased substantially as the sample size approached 40 participants, but that the benefit of additional participants diminished after that point.

At the time of the experiments, our Caucasian participants had been in the U.S. for an average of 30 years (SD = 8.5, median = 30). Among the forty Korean participants, thirty-two of them were born in South Korea and had lived in South Korea for an average of 19 years (SD = 10.02, median = 19); three were born in China, Canada, and Germany respectively and had lived in South Korea for an average of 11 years; and the other five were born in the U.S. Twenty-three of our Korean participants spoke only Korean at home, fifteen

of them spoke both Korean and English at home, and the other two spoke only English at home. All procedures were carried out in compliance with the approval of the Caltech Institutional Review Board. All participants signed a written informed consent before the study and received between $15 to $40 (depending on their travel distance) for their participation in the study. All participants completed all parts of the study and none was excluded from the analysis.

Stimuli

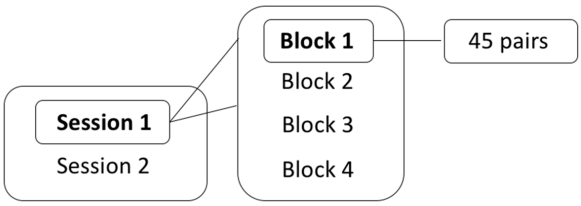
Stimuli were headshot photographs of real political candidates who ran in U.S Congressional elections, or in Korean Assembly elections. For Caucasian candidates, following the procedure in [26], we used a randomly selected set of 45 pairs of candidates (4 female pairs) from a previously establish database [1-3] (http://tlab.princeton.edu/ databases/politicians). For Korean candidates, we used the same 45 pairs of candidates (2 female pairs) as in [26]. Images were paired according to actual electoral races, with one being the winner and the other the runner-up. Only electoral races in which candidates were of the same sex and ethnicity were included. Any conspicuous background such as the capital or a national flag was removed and replaced with a gray background. All images were in black-and-white, of similar clarity, with frontal facing and centrally presented smiling faces, and were cropped to similar sizes according to the intraocular distance. When presented on the computer screen, all images had a standard size of 3.2 cm (width) x 4.5 cm (height) [1]. All materials can be accessed at https://osf.io/qx54t/?view_only=f504dcb528aa4546a2b01ee9e54f72b3.

Procedure

All experiments were carried out at the same laboratory at Caltech with the same experimenter. Participants completed two sessions of ratings on a computer: one for Caucasian candidates, and the other for Korean candidates (Fig 1a). The ordering of the two sessions was counterbalanced across participants, for both Caucasian and Korean participants. In each session, there were four blocks, each corresponding to one of the four traits: competence, open-mindedness, threat, and corruption (Fig 1a). The ordering of the four blocks was randomized for each participant. The questions on competence and threat were worded as in [38]; those on open-mindedness and corruption were worded in the same way as the competence question. In each block, participants viewed images of the 45 pairs of the political candidates (Fig 1a), and for each pair of candidates they indicated which candidate was their choice for that trait (e.g., which candidate in a pair looked more competent to hold national congressional office) (Fig 1b). The ordering of the 45 pairs of images was randomized for each participant in each block. Positions of the images were randomized in each block and counterbalanced across blocks for each participant: in each block, for half of the races the winners were positioned on the right-hand side and for the other half they were positioned on the left-hand side; the winner of a pair appeared on one side in two of the blocks (first and third blocks) and the other side in the other two blocks.

After completing each session, participants were asked whether they recognized any of the candidates. If a participant recognized any of the candidates in a pair, his/her responses for this pair of candidates were excluded from further analysis. After completing both sessions, participants completed a paper-and-pencil survey on demographic characteristics, values, and political attitudes. All data files and analysis codes can be accessed at https://osf.io/qx54t/?view_only=f504dcb528aa4546a2b01ee9e54f72b3.

a



b

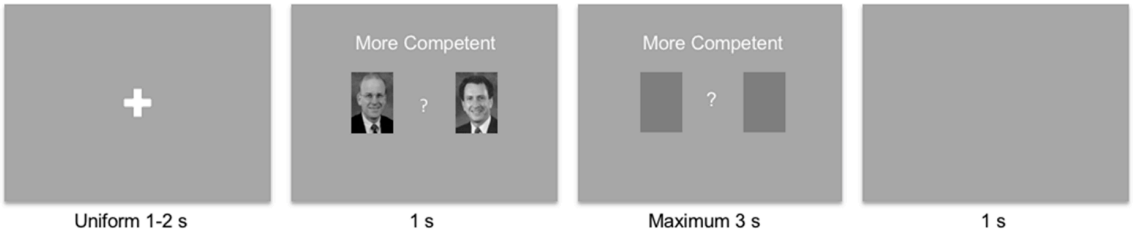


Fig 1. Experiment Procedure and Image Display Screens. (a) The schematic diagram of the full experiment. (b) Example of image display screens in the competence evaluation block. At the beginning of each block, there were instructions on the screen indicating which trait the participant was asked to evaluate. Then for each pair of candidates, participants first focused on the cross of the fixation screen that lasted for 1-2 seconds; then the images of the pair of candidates were up for 1 second; participants could make a decision as soon as the images appeared; after the images disappeared, participants had a maximum of 3 seconds to enter their choice. As soon as a valid key was pressed (i.e., press "A" if their choice was the candidate on the left and press "L" if their choice was the candidate on the right), a grey screen was up for a 1 second inter-stimulus interval.

2.3 Results

Reliability of Face-based Trait Inferences across Subjects

First, we determined the reliability of our participants' trait inferences. For each trait, we calculated the intraclass correlation coefficients (ICCs) with responses from all participants (across both cultures) for U.S. and Korean candidates respectively (using the R function ICC

(type = 'ICC2k')). As expected, across all traits and for both cultures of candidates, the ICCs were high, ranging from 0.78 to 0.87 (all p-value < 0.001), similar to those reported in [26]. These results implied that most of the variance in our participants' candidate choices was explained by the variance across the candidate pairs instead of the variance among participants. Thus, in line with previous reports [21-22, 26, 45-46], we found high consensus on face-based trait inferences across participants from both cultures.

Consistency in Face-based Inferences across Traits

Next, we determined the degree to which our participants made consistent inferences of traits. Given the high consensus on face-based trait inferences across participants, we analyzed the consistency of trait inferences at the aggregate level. For each pair of faces, we calculated for the winning candidate the percentages of participants (including both Caucasian and Korean participants) who decided he/she was their choice for being more competent, more open-minded, more threatening, and more corrupt. Using these percentages as the dependent measures, we calculated Spearman correlations between inferences on each pair of traits. We found strong positive correlations between inferences of traits with the same valence (positive or negative) and strong negative correlations between those with opposite valences (Table 1). These results suggested that, at the aggregate level, our participants made consistent trait inferences for both Caucasian and Korean candidates. It is noteworthy that the correlations we observed were nearly identical in magnitude for the evaluations of Caucasian candidates and Korean candidates. Interestingly, both perceived threat and corruption were more strongly correlated with perceived open-mindedness than perceived competence.

Table 1. Spearman Correlations between Aggregate Inferences of Different Traits

	Evaluations of Caucasian Candidates			Evaluations of Korean Candidates		
	Competence	O	T	Competence	O	T
Open-minded (O)	0.62			0.63		
	[0.39, 0.79]			[0.40, 0.79]		
Threat (T)	-0.60	-0.72		-0.58	-0.66	
	[-0.77, -0.33]	[-0.85, -0.49]		[-0.74, -0.35]	[-0.81, -0.46]	
Corruption	-0.54	-0.63	0.69	-0.60	-0.63	0.85
	[-0.74, -0.22]	[-0.76, -0.41]	[0.44, 0.83]	[-0.75, -0.37]	[-0.78, -0.39]	[0.71, 0.93]

All p-value < 0.001. 95% Confidence Intervals were presented in [].

Associations between Face-based Trait Inferences and Election Outcomes in the U.S. and Korea

Our main aim was to investigate whether face-based inferences about a range of traits about candidates were associated with which candidates won or lost in U.S. and Korean elections. We thus compared our participants' face-based trait inferences against real election outcomes. First, we looked at the data at the individual level. For each participant, we calculated the percentages of electoral races in which the candidate who was perceived as more competent, more open-minded, less threatening, and less corrupt, won the race. (Associations such as these percentages are often called "predictions" in the literature [1] even though they are fundamentally correlational and not causal in nature; to avoid confusion, we will generally use the terms "correlation" or "association".) Then, for each trait, we calculated the number of participants whose inferences agreed with the outcomes of more U.S. than Korean electoral races, and the number of participants whose inferences agreed with the outcomes of more Korean than U.S. electoral races (Fig 2). We found that the agreement between competence inferences and election outcomes were similar for U.S.

and Korean elections. On the other hand, for the majority of the participants, their inferences of open-mindedness, threat, and corruption agreed with the outcomes of more Korean electoral races than U.S. electoral races.

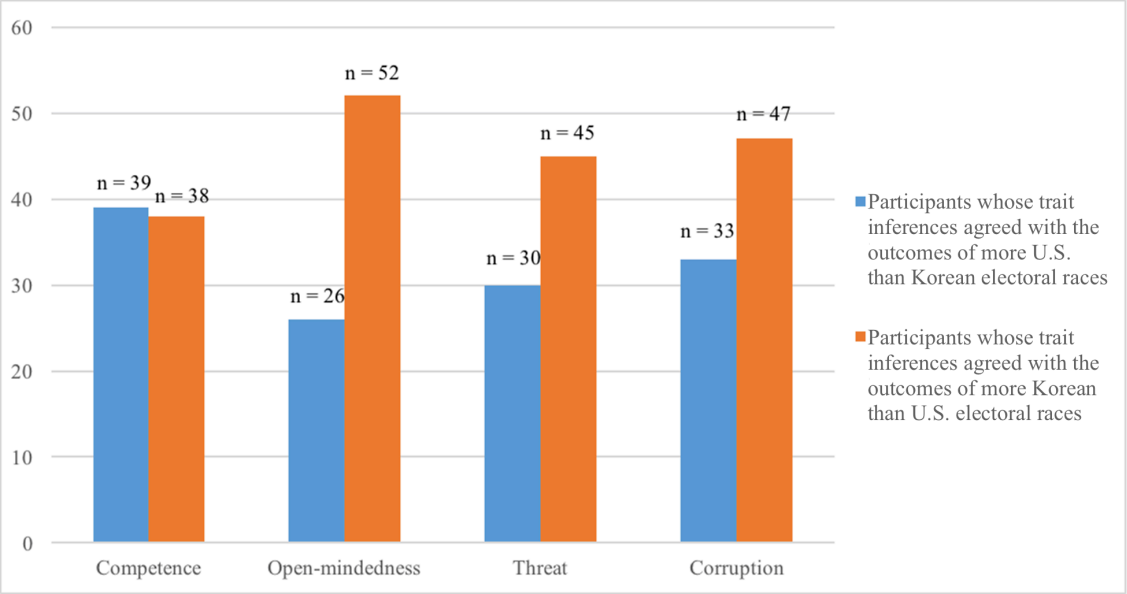


Fig 2. The Number of Participants whose Trait Inferences Agreed with the Outcomes of More Electoral Races in One Country Than the Other. The blue histogram represents the numbers of participants whose trait inferences agreed with the outcomes of more U.S. than Korean elections. The orange histogram represents the numbers of participants whose trait inferences agreed with the outcomes of more Korean than U.S. elections. For brevity, the category U.S. = Korean was omitted from the graph. All participants (N = 80).

Next, we looked at the data at the group level. We averaged these percentages of agreement over all participants (N = 80), Caucasian participants (N = 40), and Korean participants (N = 40). To see whether the agreement between inferences of a trait and election outcomes was better than chance, we performed one-sided t-tests on the percentages of agreement against 50%. To see whether the association between trait inferences and election

outcomes was stronger in one country than the other, we performed two-sided t-tests on the percentages of agreement across the two countries.

Competence

Candidates who were perceived as more competent by our participants in the lab won in more than 50% of the electoral races in both the U.S. and Korea (Table 2, columns a and b). We reproduced the results reported in the initial study [1] that Caucasian participants' judgments of competence were associated with winners in U.S. elections. Though [1] recruited college students as participants and our participants were from the general public, the average percentage of agreement we found was similar to those reported in [1] (SOM): (M = 59%, SD = 7%) for 2000 and 2002 U.S. Senate races and (M = 53%, SD = 10%) for 2004 races.

Perceived competence was associated with the outcomes of similar percentages of electoral races in the U.S. and Korea. Two-sided t-tests showed no significant difference in how well perceived competence was associated with the winning candidates in U.S. and Korean elections (Table 2, column c).

Table 2. Associations between Real Election Outcomes and Face-based Inferences of Competence

	Average Agreement		Cross-country Comparison
	U.S. Election[a]	Korean Election[b]	U.S – Korean[c]
All participants (N = 80)	**54.60%**	**54.15%**	0.45%
SD	8.46%	7.50%	t (79) = 0.38
95% CI	[53.03%, Inf)	[52.76%, Inf)	[-1.89%, 2.79%]
Caucasian participants (N = 40)	**55.33%**	**55.46%**	-0.13%

SD	8.87%	7.45%	t (39) = -0.08
95% CI	[52.97%, Inf)	[53.47%, Inf)	[-3.38%, 3.14%]
Korean participants (N = 40)	**53.87%**	**52.85%**	1.02%
SD	8.08%	7.41%	t (39) = 0.59
95% CI	[51.71%, Inf)	[50.88%, Inf)	[-2.48%, 4.51%]

[a]Average agreement between U.S. election outcomes and face-based inferences of competence, and its one-sided t-test against chance level 50%. [b]Average agreement between Korean election outcomes and face-based inferences of competence, and its one-sided t-test against chance level 50%. [c]Two-sided t-tests on the average agreement across U.S. and Korean elections.

Open-mindedness

Candidates who were perceived as more open-minded by our participants in the lab won in more than 50% of the Korean electoral races, but this association was not significant for U.S. elections (Table 3, columns a and b). Perceived open-mindedness correlated with Korean election outcomes more strongly than U.S. election outcomes (Table 3, column c).

Table 3. Associations between Real Election Outcomes and Face-based Inferences of Open-mindedness

	Average Agreement		Cross-country Comparison
	U.S. Election[a]	Korean Election[b]	U.S – Korean[c]
All participants (N = 80)	49.47%	**55.46%**	**-5.99%**
SD	8.53%	9.47%	t (79) = -3.99
95% CI	[47.89%, Inf)	[53.70%, Inf)	[-8.98%, -3.00%]
Caucasian participants (N = 40)	49.96%	**56.72%**	**-6.76%**
SD	9.41%	9.99%	t (39) = -2.81

95% CI	[47.45%, Inf)	[54.06%, Inf)	[-11.62%, -1.90%]
Korean participants (N = 40)	48.99%	**54.21%**	**-5.22%**
SD	7.63%	8.87%	t (39) = -2.86
95% CI	[46.95%, Inf)	[51.85%, Inf)	[-8.92%, -1.53%]

[a]Average agreement between U.S. election outcomes and face-based inferences of open-mindedness, and its one-sided t-test against chance level 50%. [b]Average agreement between Korean election outcomes and face-based inferences of open-mindedness, and its one-sided t-test against chance level 50%. [c]Two-sided t-tests on the average agreement across U.S. and Korean elections.

Threat

Candidates who were perceived as more threatening by our participants in the lab lost in more than 50% of the electoral races in both the U.S. and Korea, but these associations were statistically significant for only Korean elections, and not U.S. elections (Table 4, columns a and b). The average agreement (averaged over all participants and Caucasian participants) for U.S. elections significantly differed from that for Korean elections (Table 4, column c).

Table 4. Associations between Real Election Outcomes and Face-based Inferences of Threat

	Average Agreement		Cross-country Comparison
	U.S. Election[a]	Korean Election[b]	U.S – Korean[c]
All participants (N = 80)	51.50%	**54.43%**	**-2.93%**
SD	7.89%	7.38%	t (79) = -2.33
95% CI	[50.03%, Inf)	[53.05%, Inf)	[-5.43%, -0.43%]
Caucasian participants (N = 40)	51.09%	**55.89%**	**-4.80%**
SD	7.92%	7.56%	t (39) = -2.51

95% CI	[48.98%, Inf)	[53.87%, Inf)	[-8.67%, -0.93%]
Korean participants (N = 40)	51.91%	**52.97%**	-1.06%
SD	7.93%	6.97%	t (39) = -0.66
95% CI	[49.80%, Inf)	[51.11%, Inf)	[-4.29%, 2.17%]

[a]Average agreement between U.S. election outcomes and face-based inferences of threat, and its one-sided t-test against chance level 50%. [b]Average agreement between Korean election outcomes and face-based inferences of threat, and its one-sided t-test against chance level 50%. [c]Two-sided t-tests on the average agreement across U.S. and Korean elections.

Corruption

Candidates who were perceived as more corrupt by our participants in the lab lost in more than 50% of the electoral races in Korea, but this association was not significant for U.S. elections (Table 5, columns a and b). The average agreement (averaged over all participants and Caucasian participants) for U.S. elections significantly differed from that for Korean elections (Table 5, column c).

Table 5. Associations between Real Election Outcomes and Face-based Inferences of Corruption

	Average Agreement		Cross-country Comparison
	U.S. Election[a]	Korean Election[b]	U.S – Korean[c]
All participants (N = 80)	49.18%	**52.21%**	**-3.03%**
SD	9.47%	8.43%	t (79) = -2.07
95% CI	[47.42%, Inf)	[50.64%, Inf)	[-5.94%, -0.11%]
Caucasian participants (N = 40)	47.50%	**52.46%**	**-4.96%**
SD	10.28%	8.72%	t (39) = -2.29

95% CI	[44.76%, Inf)	[50.14%, Inf)	[-9.33%, -0.59%]
Korean participants (N = 40)	50.86%	51.95%	-1.09%
SD	8.38%	8.24%	t (39) = -0.56
95% CI	[48.62%, Inf)	[49.76%, Inf)	[-5.05%, 2.86%]

[a]Average agreement between U.S. election outcomes and face-based inferences of corruption, and its one-sided t-test against chance level 50%. [b]Average agreement between Korean election outcomes and face-based inferences of corruption, and its one-sided t-test against chance level 50%. [c]Two-sided t-tests on the average agreement across U.S. and Korean elections.

Response-Time Mediates the Associations between Face-based Trait Inferences and Real Election Outcomes

We investigated how response-times might be related to the above associations between face-based trait inferences and real election outcomes. We had collected a large number of individual observations (nTrial = 28540) across all participants, candidate pairs, and traits, excluding missing data, data for recognized candidates, and seven trials with response times less than 100 milliseconds (the minimum time needed for visual exploration of the faces [3]). The average response time across all trials was 1.23 seconds (SD = 0.44s). In line with prior literature, the distribution of our participants' response times was similar to the ex-Gaussian distribution (Fig 3). Interestingly, when the percentages of agreement were binned over trials within specific response-time intervals, we found a negative correlation (*rho* = -0.828, 95% CI = [-0.954, -0.453], $p = 0.002$) between response times and agreement percentages (Fig 3).

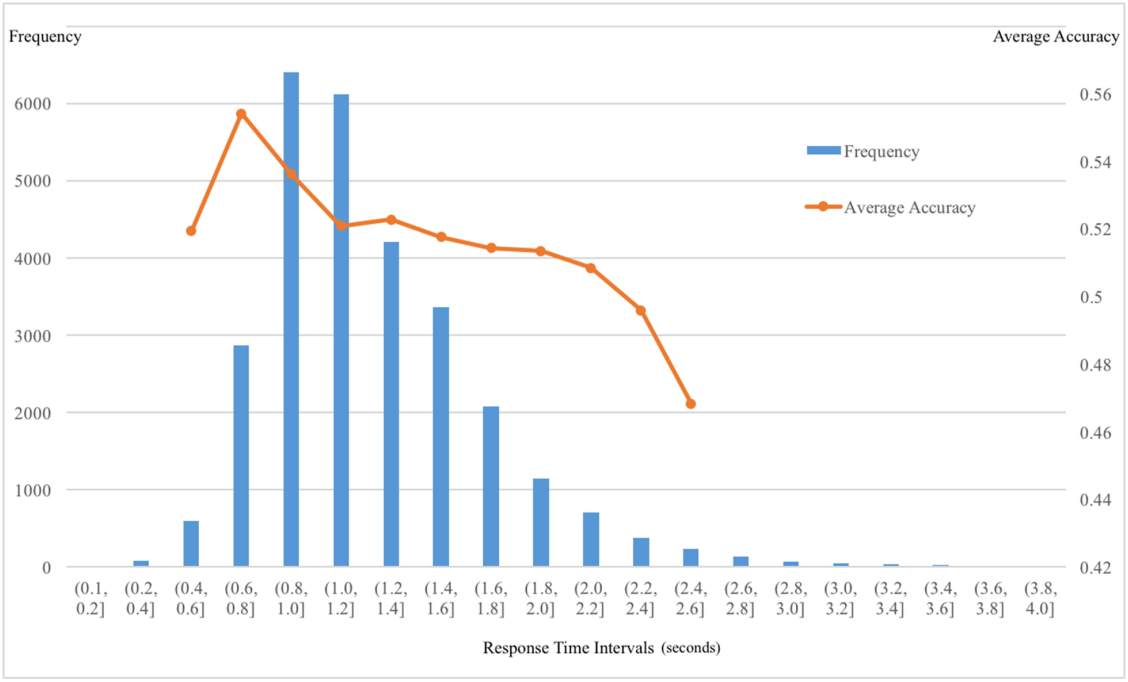


Fig 3. Distribution of Response Times and Average Agreement. The histogram represents the distribution of response times over all trials (n = 28540) across all participants, candidate pairs, and traits, excluding missing data, data for recognized candidates, and seven trials with response times less than 100 milliseconds. The line represents the average agreement over trials with response times within the given interval, omitting those for response-time intervals with less than 200 trials.

To further test the effect of response-time on the agreement between face-based trait inferences and election outcomes, we regressed the binary agreements on log-transformed response times in a logit model (Table 6, Model 1). We found that the shorter response times a participant used to make face-based trait inferences, the more likely his/her trait inferences agreed with the real election outcomes. We also controlled for candidates' cultures (Table 6, Model 2), traits (Table 6, Model 3), participants' cultures, and all other individual characteristics (Table 6, Model 4). We found a significant negative relation between time

and election agreement in all the models, even in those with extensive covariates (Table 6, Model 5; see S2 Table for the complete list of covariates). To account for the correlated errors among responses made by the same participant and those for the same candidate pair, we also clustered the standard errors at individual and image levels (S2 Table, Model 5a). Note that none of the interaction terms had a significant effect, which suggested the negative association between response-time and agreement was invariant of candidates' cultures, participants' cultures, and the traits being evaluated.

Table 6. The Effect of Response Time on the Association between Face-based Trait Inferences and Real Election Outcomes

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Log Time	**-0.112 ****	**-0.093 .**	**-0.138 ***	**-0.215 *****	**-0.229 ***
	(0.035)	(0.050)	(0.067)	(0.056)	(0.091)
Candidate Culture (1 = Korean)		**0.117 *****			**0.118 *****
		(0.026)			(0.035)
Candidate Culture * Log Time		**-0.026**			**-0.026**
		(0.069)			(0.073)
Competence			**0.135 *****		**0.120 ****
			(0.037)		(0.037)
Open-mindedness			**0.078 ***		**0.092 ***
			(0.037)		(0.037)
Threat			**0.074 ***		**0.078 ***
			(0.037)		(0.037)
Competence * Log Time			**0.073**		**0.092**
			(0.096)		(0.105)

Open-mindedness * Log Time			**-0.060**		**-0.065**
			(0.096)		(0.101)
Threat * Log Time			**0.119**		**0.157**
			(0.095)		(0.100)
Participant Culture (1 = Korean)				**-0.048**	**-0.045**
				(0.035)	(0.035)
Participant Culture * Log Time				**0.060**	**0.052**
				(0.050)	(0.050)
Gender (1 = Female)				**0.003**	**0.002**
				(0.027)	(0.027)
Age				**-0.003**	**-0.003**
				(0.003)	(0.003)
Education				**0.030 ***	**0.029 ***
				(0.012)	(0.012)
Years in U.S.				**0.005 ***	**0.005 ***
				(0.002)	(0.002)
Political Participation: Vote				**-0.108 *****	**-0.107 *****
				(0.031)	(0.031)
Liberal-Conservative Placement				**0.022 .**	**0.022 .**
				(0.013)	(0.013)
Collectivism Score				**0.278 ***	**0.266 ***
				(0.110)	(0.110)
Goodness of Fit: C-index	0.513	0.521	0.520	0.526	0.536

[1] Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' [2] In model 3, corruption was the reference trait. [3] In model 4 and 5, some insignificant individual characteristics were not presented in the table because of limited space. For the complete list of variables, please refer to (S2 Table).

2.4 Discussion

Summary of Results

We reproduced the previously reported finding that candidates perceived as more competent by Caucasian participants were associated with winners in U.S. elections [1]. This finding has been reported in numerous studies, most of which recruited students as participants [e.g., 1-3, 7, 18, 24-26, 38-39, 47-49]. Although there is concern that students may be a poor subject population for studying political decision-making [50], there is also evidence suggesting that subjects of different age groups agree on face-based judgments of competence [12]. In the present study, we recruited participants from the general public and reproduced the basic result in [1]. Moreover, the mean and variance of the percentage of agreement we found were similar to those reported in studies with student samples. Our results strengthen the external validity of the primary finding in this literature.

Contrary to [26], however, we found that inferences of competence, made by participants from both cultures, were about equally and strongly correlated with U.S. and Korean election outcomes. This discrepancy is not likely to be due to the differences in stimuli. We used the same set of Korean candidate images, and followed the same procedures in selecting Caucasian candidate images, as in [26]. Instead, the confounding order effect in [26] and the differences in subject pools between the two studies might have led to the discrepancies in findings. While the ordering of image groups (2 candidate cultures) was randomized in our study for both Caucasian and Korean participants, all Korean participants in [26] evaluated U.S. candidates first. While the Korean participants in our study had lived in the U.S. for at least six months, the Korean participants in [26] were in Korea. It is worth

noting that, in our study, how long the Korean participants had lived in the U.S. did not affect the strength of the trait-election association. We found that competence inferences made by "Long-time Koreans" (who had lived in the U.S. for a longer time than the median, 7.5 years) and "New Koreans" (who had lived in the U.S. for a shorter time than the median) were similarly associated with election outcomes (in both countries): (M_{LT} = 53.83%, M_{new} = 53.90%, d = -0.07%, t(37) = -0.03, 95% CI = [-5.32%, 5.17%]) for U.S. elections, and (M_{LT} = 54.35%, M_{new} = 51.35%, d = 3.00%, t(36) = 1.29, 95% CI = [-1.71%, 7.72%]) for Korean elections. We thus believe that the discrepancy between our findings (of cultural similarity in the association between competence judgments and election outcomes) and the findings of [26] (of cultural differences for the same association) may be traced primarily to order effects in [26].

We found that the specific traits that were most strongly associated with real election outcomes differed between the two countries: while perceived competence (by participants from both cultures) correlated with winning candidates in both countries as just noted, perceived open-mindedness and threat (by participants from both cultures) were associated with winning and losing candidates (respectively) in Korean elections only. One possible explanation for why perceived open-mindedness was associated with Korean election outcomes could be that the Asian transition from more closed to more open societies, and their adaptation to globalization, have encouraged voters to favor more reform-oriented and open-minded political leaders [51]. However, unlike [38, 39], the associations we found between perceived threat and U.S. election outcomes were not significant (though the average percentages of agreement were slightly above chance). This discrepancy is unlikely to be due to differences in stimuli, question wordings, or experimental procedures: our

stimuli for Caucasian candidates were randomly selected from the same face database as in [39], our threat evaluation question was worded identically, and our image presentation procedure was also identical to [39]. One possible explanation for the discrepancy might be that student samples (in [39]) and general public samples (in our experiments) differ in how they perceive threat from faces. It will be important for future studies to investigate how student samples and non-student samples might differ in making face-based inferences across a broader profile of negative traits, and how such judgments may depend also on the personality of the viewer.

Implications for Causality

While our study is fundamentally correlational in nature, the findings nonetheless have implications for causal hypotheses. Several studies have investigated whether candidate appearances causally influence voter decisions [30, 44, 52-53]. If voters take visual cues from candidates' physical appearances when they decide which candidate to vote for, then one would expect that the impact of appearances is greater on those who are exposed to more visual images of the candidates. One of the studies [52] tested this hypothesis on a combined dataset with individual-voter-level data about vote intent, political knowledge, and TV exposure, and candidate-level data about the ratings of their appearances. They found that the effect of candidate appearances was more pronounced among those who had high TV exposure but knew little about the candidates. Another of these studies [53] tested the causal hypothesis by conducting two internet polls in which registered voters intending to vote were randomly assigned to receive standard ballots or ballots with candidate photos. They found that better-looking candidates experienced greater success in the ballots with their photos than the standard ballots, and that this effect was stronger among low-knowledge voters.

On the other hand, the cultural differences we found provide a new perspective on testing the causal relationship between candidate appearances and election outcomes. If voters evaluate candidates on the traits they value and take visual cues from faces for these evaluations, then one would expect that the specific traits that most strongly associate with election outcomes would differ across cultures because people from different cultures value different traits of their leaders [54, 55]. In our study, almost all cultural effects were driven by the culture of the politicians, not the culture of the participants, which suggests that the differential effect of various traits on election results might arise from how those traits are valued in the respective cultures. To provide causal evidence, future studies could investigate whether open-mindedness and threat have stronger impacts on impression formation and leader evaluation in Korea (or Asian countries) than the U.S. (or Caucasian cultural countries).

Implications for Political Behavior

The cultural differences we found also have implications for the study of political behavior. In the classic vote choice model, major considerations were given to social determinants, party identification, and political issues. Studies trying to measure the effects of candidate traits on election outcomes found conflicting results: there was evidence that assessments of candidate traits influenced individual vote choice [56-58], with some arguing that the effects of candidate traits might be mediated by uncertainty and information [59, 60], while others asserted that the net effects of candidate traits might be negligible [61, 62]. Our findings have demonstrated that candidate traits have significant effects on elections and should be included in the classic vote choice model.

Implications for Cognitive Psychology

Counter to the usual speed-accuracy trade-off, we found that the shorter the response times a participant took to make face-based trait inferences, the more strongly his/her trait inferences correlated with election outcomes. This finding provides new insights into the higher cognitive processes that might be involved in face-based impression formation. Prior to our study, some [2, 63] have investigated the effects of response-time on the association between trait inferences and election outcomes. By manipulating image exposure time and the response deadline procedure, those studies found that increasing image exposure time after 100 milliseconds did not strengthen the association, and instructing subjects to deliberate in fact weakened the association. However, based on these prior findings, it is not straightforward to conclude that under the same image exposure time and response deadline condition, shorter response times should result in stronger associations, as we found in our study. Moreover, we found negative correlations between response-times and the trait-election associations regardless of candidates' cultures, participants' cultures, or the types of traits being evaluated. Faster trait judgments always produced stronger associations.

We suggest two possible explanations for this effect of response-time, which require further investigation. First, the quicker a participant is to make a choice between a pair of candidates, the more likely it is that these two candidates look different, making it easier for the participant to decide which one fits the trait better. On the other hand, taking a longer time to make a choice between two candidates suggests greater uncertainty and difficulty, and therefore the decision tends to be less accurate. Thus, short response times may be correlated with stronger trait-election associations simply because they are derivative to those judgments about pairs of politicians that are also the easiest to make.

A second, and not mutually exclusive, possibility is that evaluating candidates on certain traits by real-world voters may engage mostly "system 1" processes (a type of cognitive process that is quick, automatic, and effortless [64]). That is, when voters actually vote for candidates, they may well be incorporating trait judgments about the candidates into their choices—but such judgments at the time of voting would likely be implicit, automatic processes more aligned with "system 1". Participants in our experiment, on the other hand, might exhibit a range of processes when making their trait judgments, as reflected in the range of reaction times that they produced. Some of those judgments – the ones with short reaction times – could plausibly be in line with "system 1" processes; whereas, other judgments – the ones with long reaction times – could plausibly reflect "system 2" processes (another distinct type of higher cognitive process that is slow and requires effort [64]), which perhaps even to correct the snap judgments made by system 1. Those trials in the lab with short reaction times might then correspond more closely to the evaluative processing in voters which influences their actual choices (both are "system 1"), and hence show the strongest association with election outcomes. While this second hypothesis is of course very speculative at this stage, it makes predictions about the type of psychological processes that could actually influence voters at the time that they make their election choices, predictions that could be tested in future studies.

Other Mediational Effects

It is also interesting that candidate appearances might have stronger effects on some voters than on others. Recent studies [30, 52-53] have investigated how access to information influences the impact of candidate appearances on voter decisions. These studies found that voters with less political information relied more on candidate appearances in their decision-

making. We found that inferences made by participants who had lower levels of political participation were more strongly associated with real election outcomes (Table 6). As suggested in our results and [25], individualistic-collectivist orientations might mediate the association between candidate appearances and voter decisions as well. Moreover, political ideology might be yet another contributing factor. One study [29] found that candidates facing conservative electorates benefited from looking more stereotypically Republican, while no relationship between political facial stereotypes and voting was found for liberal electorates. Another study [65] suggested that voters on the right were more responsive to beautiful candidates than voters on the left. In our study, inferences made by more conservative participants were more strongly associated with election outcomes, but this effect of political ideology became insignificant when the correlated errors were adjusted.

Our last point is that some of the images we used are more than a decade old (e.g., some images were of candidates from the 2000 U.S. Senate elections). The development of social media and image processing technology, and the awareness of the association between candidate appearances and election outcomes in the past decade, may have changed the relationship between attribute judgments and election outcomes. It will be important to investigate how the relationships that have been reported to date may change over time.

2.5 Supplementary Information

S1 Table. Subject pool demographic statistics.

	Caucasian Participants (N = 40, 20 male)		Korean Participants (N = 40, 20 male)	
	M	SD	M	SD
Age	31	6.93	29	6.42
Education[a]	Associates' Degree	-	Bachelor's Degree	-
Wilson Patterson Scale[b]	0.30	0.13	0.45	0.10
Ideology Placement[c]	2.63	1.37	3.58	1.26
Voted in Election[d]	65.79%	-	32.50%	-
Days/week discussed politics[e]	3.1	1.79	1.23	1.14
Political Campaign[f]	68.42%	-	37.50%	-
Political Knowledge Score[g]	0.70	0.19	0.57	0.20
Individualism Score[h]	0.71	0.14	0.64	0.13
Collectivism Score	0.67	0.17	0.66	0.12

[a]Participants indicated the highest level of education they have completed. The medians were presented in the table.

[b]Participants indicated on a 5-point scale how they felt about each topic in the 16-item Wilson Paterson Scale [66]. Response options were strongly agree, agree, uncertain, disagree, and strongly disagree. Responses to conservatism oriented items were scored from 5 to 1; responses to other items were reversed, thus scored from 1 to 5. Scores for the 16 items were averaged for each participant.

[c]Participants indicated their ideology on a 7-point liberal-conservative scale. The question was from American National Election Studies (ANES) Pre 2012. Scores were normalized to [0,1].

[d]Participants indicated whether or not {0,1} they had voted in the last Presidential Election they were eligible to vote, no matter it was in the U.S. or Korea. The question was a variation of Current Population Survey (CPS) Nov 2012, PES1.

[e]Participants indicated during a typical week how many days {0,…,7} they discussed politics with their family or friends. The question was from ANES post 2008.

[f]Participants indicated during the campaign whether they talked to people about who they should vote for or against. The question was from ANES post 2004.

[g]Participants were asked seven questions on their political knowledge. Questions varied in their difficulty and their relevance to U.S. and Korea (e.g., a question asking about United Nations meeting in New York last September, and a question asking what political office Ban Ki-moon currently holds). Scores for each question were weighted on the reversed percentages of participants getting the correct answer; thus, the more difficult the higher the weight. Weighted average scores were normalized to [0,1].

[h]Individualism and collectivism scores were calculated based on participants' agreement/disagreement on 16 statements [67].

S2 Table. The Effect of Response Time on the Association between Face-based Trait Inferences and Real Election Outcomes.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 5a Adj Errors
Log Time	-0.112 **	-0.093 .	-0.138 *	-0.215 ***	-0.229 *	**-0.229 .**
	(0.035)	(0.050)	(0.067)	(0.056)	(0.091)	(0.137)
Candidate Culture (1 = Korean)		0.117 ***			0.118 ***	0.118
		(0.026)			(0.035)	(0.107)
Candidate Culture * Log Time		-0.026			-0.026	-0.026
		(0.069)			(0.073)	(0.105)
Competence			0.135 ***		0.120 **	**0.120 ***
			(0.037)		(0.037)	(0.053)
Open-mindedness			0.078 *		0.092 *	**0.092 .**
			(0.037)		(0.037)	(0.048)
Threat			0.074 *		0.078 *	**0.078 .**
			(0.037)		(0.037)	(0.042)
Competence * Log Time			0.073		0.092	0.092
			(0.096)		(0.105)	(0.129)
Open-mindedness * Log Time			-0.060		-0.065	-0.065
			(0.096)		(0.101)	(0.119)
Threat * Log Time			0.119		0.157	0.157
			(0.095)		(0.100)	(0.117)
Participant Culture (1 = Korean)				-0.048	-0.045	-0.045

				(0.035)	(0.035)	(0.047)
Participant Culture * Log Time				0.060	0.052	0.052
				(0.050)	(0.050)	(0.061)
Gender (1 = Female)				0.003	0.002	0.002
				(0.027)	(0.027)	(0.030)
Age				-0.003	-0.003	-0.003
				(0.003)	(0.003)	(0.003)
Education				0.030 *	0.029 *	**0.029 ***
				(0.012)	(0.012)	(0.013)
Years in U.S.				0.005 *	0.005 *	0.005
				(0.002)	(0.002)	(0.003)
Political Participation: Vote				-0.108 ***	-0.107 ***	**-0.107 ****
				(0.031)	(0.031)	(0.038)
Political Participation: Talk Politics				0.001	0.001	0.001
				(0.009)	(0.009)	(0.009)
Political Participation: Campaign				-0.010	-0.010	-0.010
				(0.027)	(0.027)	(0.026)
Wilson Patterson Scale				0.142	0.149	0.149
				(0.152)	(0.152)	(0.192)
Liberal-Conservative Placement				0.022 .	0.022 .	0.022
				(0.013)	(0.013)	(0.020)
Political Knowledge Score				-0.117 .	-0.110	-0.110
				(0.070)	(0.070)	(0.099)
Individualism Score				-0.010	-0.012	-0.012
				(0.105)	(0.105)	(0.112)
Collectivism Score				0.278 *	0.266 *	**0.266 .**
				(0.110)	(0.110)	(0.139)
Goodness of Fit: C-index	0.513	0.521	0.520	0.526	0.536	0.536

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

2.6 References

1. Todorov A, Mandisodza AN, Goren A, Hall CC. Inferences of competence from faces predict election outcomes. *Science*. 2005 Jun 10; 308(5728):1623–6. https://doi.org/10.1126/science.1110589 PMID: 15947187
2. Ballew CC, Todorov A. Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences*. 2007 Nov 13; 104(46):17948–53. https://doi.org/10. 1073/pnas.0705435104 PMID: 17959769
3. Olivola CY, Todorov A. Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*. 2010 Jun 1; 34(2):83–110.
4. Carpinella CM, Johnson KL. Visual Political Communication: The Impact of Facial Cues from Social Constituencies to Personal Pocketbooks. *Social and Personality Psychology Compass*. 2016 May 1; 10 (5):281–97.
5. Martin DS. Person perception and real-life electoral behaviour. *Australian Journal of Psychology*. 1978 Dec 1; 30(3):255–62.
6. Banducci SA, Karp JA, Thrasher M, Rallings C. Ballot photographs as cues in low-information elections. *Political Psychology*. 2008 Dec 1; 29(6):903–17. https://doi.org/10.1111/j.1467-9221.2008.00672.x
7. Mattes K, Milazzo C. Pretty faces, marginal races: Predicting election outcomes using trait assessments of British parliamentary candidates. *Electoral Studies*. 2014 Jun 30; 34:177–89. https://doi.org/ 10.1016/j.electstud.2013.11.004.

8. Milazzo C, Mattes K. Looking good for election day: Does attractiveness predict electoral success in Britain?. *The British Journal of Politics and International Relations*. 2016 Feb; 18(1):161–78.
9. Little AC, Burriss RP, Jones BC, Roberts SC. Facial appearance affects voting decisions. *Evolution and Human Behavior*. 2007 Jan 31; 28(1):18–27.
10. Rosar U, Klein M, Beckers T. The frog pond beauty contest: Physical attractiveness and electoral success of the constituency candidates at the North Rhine-Westphalia state election of 2005. *European Journal of Political Research*. 2008 Jan 1; 47(1):64–79.
11. Stockemer D, Praino R. Physical attractiveness, voter heuristics and electoral systems: The role of candidate attractiveness under different institutional designs. *The British Journal of Politics and International Relations*. 2017 Feb 8:1369148116687533.
12. Antonakis J, Dalgas O. Predicting elections: Child's play!. *Science*. 2009 Feb 27; 323(5918):1183-. https://doi.org/10.1126/science.1167748 PMID: 19251621.
13. Berggren N, Jordahl H, Poutvaara P. The looks of a winner: Beauty and electoral success. *Journal of Public Economics*. 2010 Feb 28; 94(1):8–15.
14. Poutvaara P, Jordahl H, Berggren N. Faces of politicians: Babyfacedness predicts inferred competence but not electoral success. *Journal of Experimental Social Psychology*. 2009 Sep 30; 45(5):1132–5.
15. Buckley F, Collins N, Reidy T. Ballot Paper Photographs and Low-Information Elections in Ireland. *Politics*, 2007; 27(3): 174–181. https://doi.org/10.1111/j.1467-9256.2007.00297.x

16. Lutz G. The electoral success of beauties and beasts. *Swiss Political Science Review*. 2010 Sep 1; 16 (3):457–80.

17. Sussman AB, Petkova K, Todorov A. Competence ratings in US predict presidential election outcomes in Bulgaria. *Journal of Experimental Social Psychology*. 2013 Jul 31; 49(4):771–5. https://doi.org/10. 1016/j.jesp.2013.02.003

18. Laustsen L. Decomposing the relationship between candidates' facial appearance and electoral success. *Political Behavior*. 2014 Dec 1; 36(4):777–91.

19. Castelli L, Carraro L, Ghitti C, Pastore M. The effects of perceived competence and sociability on electoral outcomes. *Journal of Experimental Social Psychology*. 2009 Sep 30; 45(5):1152–5.

20. King A, Leigh A. Beautiful politicians. Kyklos. 2009 Nov 1; 62(4):579–93.

21. Lawson C, Lenz GS, Baker A, Myers M. Looking like a winner: Candidate appearance and electoral success in new democracies. *World Politics*. 2010 Oct 1; 62(04):561–93.

22. Rule NO, Ambady N, Adams RB Jr, Ozono H, Nakashima S, Yoshikawa S, et al. Polling the face: prediction and consensus across cultures. *Journal of personality and social psychology*. 2010 Jan; 98(1):1. https://doi.org/10.1037/a0017673 PMID: 20053027

23. Wong SH, Zeng Y. Do inferences of competence from faces predict political selection in authoritarian regimes? Evidence from China. *Social Science Research*. 2016 Nov 16.

24. Chen FF, Jing Y, Lee JM, Bai L. Culture Matters The Looks of a Leader Are Not All the Same. *Social Psychological and Personality Science*. 2016 Apr 28:1948550616644962.

25. Chen FF, Jing Y, Lee JM. "I" value competence but "we" value social competence: The moderating role of voters' individualistic and collectivistic orientation in political elections. *Journal of Experimental Social Psychology*. 2012 Nov 30; 48(6):1350–5.

26. Na J, Kim S, Oh H, Choi I, O'Toole A. Competence judgments based on facial appearance are better predictors of American elections than of Korean elections. *Psychological science*. 2015 May 8:0956797615576489. https://doi.org/10.1177/0956797615576489 PMID: 25956912

27. Horiuchi Y, Komatsu T, Nakaya F. Should candidates smile to win elections? An application of automated face recognition technology. *Political Psychology*. 2012 Dec 1; 33(6):925–33.

28. Tsai JL, Ang JY, Blevins E, Goernandt J, Fung HH, Jiang D, et al. Leaders' smiles reflect cultural differences in ideal affect. *Emotion*. 2016 Mar; 16(2):183. https://doi.org/10.1037/emo0000133 PMID: 26751631

29. Olivola CY, Sussman AB, Tsetsos K, Kang OE, Todorov A. Republicans prefer Republican-looking leaders: Political facial stereotypes predict candidate electoral success among right-leaning voters. *Social Psychological and Personality Science*. 2012 Sep; 3(5):605–13.

30. Olivola CY, Funk F, Todorov A. Social attributions from faces bias human choices. *Trends in Cognitive Sciences*. 2014 Nov 30; 18(11):566–70. https://doi.org/10.1016/j.tics.2014.09.007 PMID: 25344029

31. Hess U, Blairy S, Kleck RE. The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation. *Journal of Nonverbal behavior*. 2000 Dec 1; 24(4):265–83.

32. Montepare JM, Dobish H. The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of Nonverbal behavior*. 2003 Dec 1; 27(4):237–54.

33. Oosterhof NN, Todorov A. Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion*. 2009 Feb; 9(1):128. https://doi.org/10.1037/a0014520 PMID: 19186926

34. Said CP, Sebe N, Todorov A. Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion*. 2009 Apr; 9(2):260. https://doi.org/10.1037/a0014681 PMID: 19348537

35. Zebrowitz LA, Kikuchi M, Fellous JM. Facial resemblance to emotions: group differences, impression effects, and race stereotypes. *Journal of personality and social psychology*. 2010 Feb; 98(2):175. https://doi.org/10.1037/a0017990 PMID: 20085393

36. Kim H, Markus HR. Deviance or uniqueness, harmony or conformity? A cultural analysis. *Journal of personality and social psychology*. 1999 Oct; 77(4):785.

37. Tsai JL, Knutson B, Fung HH. Cultural variation in affect valuation. *Journal of personality and social psychology*. 2006 Feb; 90(2):288. https://doi.org/10.1037/0022-3514.90.2.288 PMID: 16536652

38. Spezio ML, Rangel A, Alvarez RM, O'Doherty JP, Mattes K, Todorov A, et al. A neural basis for the effect of candidate appearance on election outcomes. *Social*

Cognitive and Affective Neuroscience. 2008 Dec 1; 3(4):344–52. https://doi.org/10.1093/scan/nsn040 PMID: 19015087

39. Mattes K, Spezio M, Kim H, Todorov A, Adolphs R, Alvarez RM. Predicting election outcomes from positive and negative trait assessments of candidate images. *Political Psychology*. 2010 Feb 1; 31(1):41– 58.
40. Olivola CY, Eubanks DL, Lovelace JB. The many (distinctive) faces of leadership: Inferring leadership domain from facial appearance. *The Leadership Quarterly*. 2014 Oct 31; 25(5):817–34.
41. Fiske ST, Cuddy AJ, Glick P. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*. 2007 Feb 28; 11(2):77–83. https://doi.org/10.1016/j.tics.2006.11.005 PMID: 17188552
42. Oosterhof NN, Todorov A. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*. 2008 Aug 12; 105(32):11087–92. https://doi.org/10.1073/pnas.0805664105 PMID: 18685089
43. Cogsdill EJ, Todorov AT, Spelke ES, Banaji MR. Inferring character from faces a developmental study. *Psychological science*. 2014 May 1; 25(5):1132–9. https://doi.org/10.1177/0956797614523297 PMID: 24570261
44. Todorov A, Olivola CY, Dotsch R, Mende-Siedlecki P. Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*. 2015 Jan 3; 66:519– 45. https://doi.org/10.1146/annurev-psych-113011-143831 PMID: 25196277

45. Albright L, Malloy TE, Dong Q, Kenny DA, Fang X, Winquist L, et al. Cross-cultural consensus in personality judgments. *Journal of personality and social psychology*. 1997 Mar; 72(3):558. PMID: 9120784

46. Langlois JH, Kalakanis L, Rubenstein AJ, Larson A, Hallam M, Smoot M. Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological bulletin*. 2000 May; 126(3):390. PMID: 10825783

47. Atkinson MD, Enos RD, Hill SJ. Candidate faces and election outcomes: Is the face-vote correlation caused by candidate selection. *Quarterly Journal of Political Science*. 2009 Apr 17; 4(3):229–49.

48. Spezio ML, Loesch L, Gosselin F, Mattes K, Alvarez RM. Thin-Slice Decisions Do Not Need Faces to be Predictive of Election Outcomes. *Political Psychology*. 2012 Jun 1; 33(3):331–41.

49. Chen FF, Jing Y, Lee JM. The looks of a leader: Competent and trustworthy, but not dominant. *Journal of Experimental Social Psychology*. 2014 Mar 31; 51:27–33.

50. Lau RR, Redlawsk DP. How voters decide: Information processing in election campaigns. Cambridge University Press; 2006 Jun 26.

51. Guo S. The political economy of Asian transition from communism. Ashgate Publishing, Ltd.; 2006.

52. Lenz GS, Lawson C. Looking the part: Television leads less informed citizens to vote based on candidates' appearance. *American Journal of Political Science*. 2011 Jul 1; 55(3):574–89.

53. Ahler DJ, Citrin J, Dougal MC, Lenz GS. Face value? Experimental evidence that candidate appearance influences electoral choice. *Political Behavior*. 2016:1–26.

54. Gerstner CR, Day DV. Cross-cultural comparison of leadership prototypes. *The Leadership Quarterly*. 1994 Jun 1; 5(2):121–34.

55. Ensari N, Murphy SE. Cross-cultural variations in leadership perceptions and attribution of charisma to the leader. *Organizational Behavior and Human Decision Processes*. 2003 Nov 30; 92(1):52–66.

56. Miller AH, Reisinger WM, Hesli VL. Leader Popularity and Party Development in Post-Soviet Russia. 1998

57. Newman BI. Politics in an age of manufactured images. *Journal of mental changes*. 1999; 5(2):7–26

58. Costa P, Ferreira da Silva F. The impact of voter evaluations of leaders' traits on voting behaviour: Evidence from seven European Countries. *West European Politics.* 2015 Nov 2; 38(6):1226–50.

59. Glasgow G, Alvarez RM. Uncertainty and candidate personality traits. *American Politics Quarterly*. 2000 Jan; 28(1):26–49.

60. Alvarez RM. Information and elections. University of Michigan Press; 1998.

61. Miller WE, Shanks JM. The new American voter. Cambridge, MA: Harvard University Press; 1996 Oct.

62. Bartels LM. The impact of candidate traits in American presidential elections. Leaders' personalities and the outcomes of democratic elections. 2002:44–69.

63. Willis J, Todorov A. First impressions making up your mind after a 100-ms exposure to a face. *Psychological science*. 2006 Jul 1; 17(7):592–8. https://doi.org/10.1111/j.1467-9280.2006.01750.x PMID: 16866745

64. Kahneman D. Thinking, fast and slow. Macmillan; 2011 Oct 25. 65. Berggren N, Jordahl H, Poutvaara P. The right look: Conservative politicians look better and voters reward it. *Journal of Public Economics*. 2017 Feb 28; 146:79–86.

65. Ahn WY, Kishida KT, Gu X, Lohrenz T, Harvey A, Alford JR, Smith KB, Yaffe G, Hibbing JR, Dayan P, Montague PR. Nonpolitical images evoke neural predictors of political ideology. *Current Biology*. 2014 Nov 17;24(22):2693-9. doi: 10.1016/j.cub.2014.09.050

66. Singelis TM, Triandis HC, Bhawuk DP, Gelfand MJ. Horizontal and vertical dimensions of individualism and collectivism: A theoretical and measurement refinement. *Cross-cultural research*. 1995 Aug 1;29(3):240-75. doi: 10.1177/106939719502900302.

Chapter 3

Inferring Corruptible Officials from Their Faces

While inferences of traits from unfamiliar faces prominently reveal stereotypes, some facial inferences also correlate with real-world outcomes. We investigated whether facial inferences are associated with an important real-world outcome closely linked to the face bearer's behavior: political corruption. In four preregistered studies (N = 325), participants made trait judgments of unfamiliar government officials on the basis of their photos. Relative to peers with clean records, federal and state officials convicted of political corruption (Study 1) and local officials who violated campaign finance laws (Study 2) were perceived as more corruptible, dishonest, selfish, and aggressive but similarly competent, ambitious, and masculine (Study 3). Mediation analyses and experiments in which the photos were digitally manipulated showed that participants' judgments of how corruptible an official looked were causally influenced by the face width of the stimuli (Study 4). The findings shed new light on the complex causal mechanisms linking facial appearances with social behavior.

3.1 Introduction

Faces are rich in information: They provide clues about gender, race, age, and trait attributes, which are inferred spontaneously and ubiquitously (Engell, Haxby, & Todorov, 2007; Todorov, 2017). Moreover, such inferences often guide our social behavior—for instance, we decide whom to trust on the basis of how trustworthy a face looks (Rezlescu, Duchaine, Olivola, & Chater, 2012; Van't Wout & Sanfey, 2008). Many trait judgments made by participants across generations and cultures show consensus (Cogsdill, Todorov, Spelke, & Banaji, 2014; Lin, Adolphs, & Alvarez, 2017; Rule et al., 2010). But are trait judgments from faces accurate?

Previous research has shown that trait judgments from faces can be associated with important real-world social outcomes, such as dating and mating (Olivola et al., 2014; Valentine, Li, Penke, & Perrett, 2014), earnings and fundraising (Genevsky & Knutson, 2015; Hamermesh, 2011; Ravina, 2012), science communication (Gheorghiu, Callan, & Skylark, 2017), sentencing decisions (Berry & Zebrowitz-McArthur, 1988; Blair, Judd, & Chapleau, 2004; Wilson & Rule, 2015; Zebrowitz & McDonald, 1991), and leader selection (Todorov, Mandisodza, Goren, & Hall, 2005; for reviews, see Antonakis & Eubanks, 2017; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). Yet this prior research on the association between trait judgments from faces and real-world outcomes leaves open two important questions. First, most associations have focused on prosocial outcomes (e.g., correlations between competence judgments and election success; Todorov et al., 2005). Second, most associations are plausibly driven not by the behavior of the targets whose face is being judged, but by the interests of the perceivers who are making the judgments (e.g.,

correlations between interesting-looking scientists and the perceiver's interest in their work). Here, we investigated an antisocial judgment that may offer a clearer insight into associations with the judged person's own behavior: political corruption.

Political corruption has been a major cause of regime change and an important subject of much study in political science and economics (Rose-Ackerman, 2013). The possibility that corruptibility inferences from faces might be associated with real-world measures of corruption is raised by three areas of previous research. First, theories of self-fulfilling prophecy argue that the impressions and expectations a face creates (e.g., how corruptible an official looks) influence how other people interact with the face bearer (e.g., how likely others would be to bribe the official) and that those recurrent interactions in turn shape the face bearer's behavior so as to confirm other people's impressions and expectations (Haselhuhn, Wong, & Ormiston, 2013; Jussim, 1986; Slepian & Ames, 2016). Second, analyses of sentencing decisions show that evaluations of guilt and recommendations of punishment are influenced by the defendant's facial appearance (Berry & Zebrowitz-McArthur, 1988; Blair et al., 2004; Wilson & Rule, 2015; Zebrowitz & McDonald, 1991). These findings suggest that officials who look more corruptible might be more likely to be accused, prosecuted, and convicted. Third, some studies have argued that the face contains a kernel of truth about a person's nature—such as personality and criminal inclinations (Penton-Voak, Pound, Little, & Perrett, 2006; Valla, Ceci, & Williams, 2011)—even though the diagnostic validity and the causal mechanisms remain obscure.

Given past research, we hypothesized that elected officials' corruption records would be associated with traits, such as corruptibility, inferred from their facial appearances. We examined this association in three preregistered studies, where participants made trait

inferences on the basis of the photos of unfamiliar government officials. To account for the possibility that this association might depend on the severity of corruption and the level of office, we inspected both serious violations (i.e., cases considered political corruption) and minor violations (i.e., cases meriting a fine) and included officials at different levels of government (federal, state, and local). In a fourth preregistered study, we explored which facial features might be causally mediating the impression of how corruptible an official was, using mediation analyses as well as experimental manipulations of the face stimuli. In this fourth study, we focused on metrics of facial structures—in particular, facial width (relative to facial height) because it has been reported that men with wider faces are judged as less trustworthy (Haselhuhn et al., 2013; Stirrat & Perrett, 2010), more threatening (Geniole, Denson, Dixson, Carré, & McCormick, 2015), and less than fully human (Deska, Lloyd, & Hugenberg, 2018), although it remains unknown whether facial width-to-height ratio associates with actual behavior.

We have reported all measures, all conditions, all data exclusions, and how sample sizes were determined in this article and on the Open Science Framework (https://osf.io/k4mds/). All materials, data, and analysis codes for the present research can be accessed at this link.

3.2 Study 1

Our first study focused on federal and state officials, and compared those who had clean records with those who were convicted of political corruption.

Method

Participants. This study was preregistered before data collection began (https://osf.io/mge8r/). A sample size of 100 participants was predetermined on the basis of two pilot studies—one carried out in the lab in May 2016, and the other via Amazon's Mechanical Turk (MTurk) in October 2016. The lab study included 32 participants recruited from the general public of Southern California, and the MTurk study had 18 participants. For the hypothesis that elected officials' corruption records would be associated with face-based inferences of corruptibility, the laboratory pilot study yielded an estimated effect size of 1.06, and the MTurk pilot study yielded an estimated effect size of 1.05, justifying a minimum sample size of 16 participants. Given these results and to ensure sufficient power even with dropout, we recruited 100 MTurk participants in November 2016. We selected participants who were native English speakers, located in the United States, and 18 years old or older. In addition, they had to have normal or corrected-to-normal vision, an educational attainment of high school or above, a good MTurk participation history (a human-intelligence-task, or HIT, approval rate $\geq$ 95% and $\geq$ 1,000 HITs approved), and no prior participation in our pilot studies.

Eighteen individuals were excluded in total, 2 for not being native English speakers, 6 for pressing the same response key for all trials in a block, and 10 for failing to input valid responses for more than 10% of the trials in a block (responses were considered not valid if missing or entered within 100 ms—the minimum time needed for visual exploration of the face; Olivola & Todorov, 2010). After exclusion, there were 82 participants in our final sample (42 female; age: M = 39 years, SD = 12; 84% White, 10% Black, 5% Asian).

Stimuli. Stimuli were photos of 72 real elected officials. All were Caucasian males who have held federal or state legislative offices in the United States. Photos were official

headshots obtained from government websites and personal campaign websites (63%), news articles (23%), and Wikipedia (14%). All photos were converted to gray-scale images on a plain gray background and cropped to a uniform size. All faces were frontal, smiling, in clear focus, and centered in the middle of the image.

Among the 72 officials, half were convicted of political corruption (corrupt officials), and the other half had clean records (noncorrupt officials). The corrupt officials were from two Wikipedia data sets (https://en.wikipedia.org/wiki/List_of_American_state_and_local_politicians_convicted_of_crimes; https://en.wikipedia.org/wiki/List_of_American_federal_politicians_convicted_of_crimes). To reduce sources of variability, we included only officials who were Caucasian, were male, held federal or state legislative offices, and were convicted between 2001 and 2016 of political corruption conducted while in office (bribery, money laundering, embezzlement, mail fraud, wire fraud, tax fraud, conflict of interest, misusing funds, misusing office, or falsifying records). In addition, age information for these officials had to be publicly available, as did frontal photographs of acceptable clarity in which the official was smiling.

All photographs had been taken while officials were in office. Most photos of the corrupt officials had a known creation date, and we confirmed that the photos were taken before their conviction (72%); for the rest of the photos (28%), the creation date was unknown (analyses were also per- formed when excluding data for these stimuli; the pattern of results did not change). The noncorrupt officials were randomly matched from the list of incumbents who had clean records, were holding the same office in the same state, and were of the same gender, the same race, and similar age (±12 years) as the corrupt officials during the period of their misconduct. For instance, if the stimuli contained a

Caucasian male corrupt official who was a member of the Arizona House of Representatives during his misconduct at the age of 55, then a noncorrupt official would be randomly selected from our available stimulus set from the list of Arizona House of Representatives incumbents who had a clean record and who was a Caucasian male between the ages of 43 and 67.

Procedures. Participants were not informed of the purpose of the study or the sampling of the stimuli. In particular, they were not given any information about the percentage of politicians in our stimulus set who might be corrupt in real life. They were told only that they would view a series of politician photos and that they should judge how corruptible, dishonest, selfish, trustworthy, and generous these politicians looked to them (experiment instructions are available at https://osf.io/ k4mds/). Participants completed five blocks of experiments, with each block corresponding to judging one trait for all faces. The ordering of the faces within each block as well as the ordering of the blocks were randomized.

Each block started with an instruction screen that specified the trait to be judged (e.g., corruptibility). Participants were instructed to make their decisions as quickly and precisely as possible. Six practice trials familiarized participants with the task. Participants viewed photos of officials one at a time in randomized order and made judgments. Each trial began with a fixation cross, followed by the photo (1 s) with a 5-point Likert scale below it. Scales were anchored with bipolar adjectives (Fig. 1). Participants could make a decision as soon as the photo appeared, and within 4 s after the photo disappeared. The orientation of the scale was randomized across blocks, and scores were reverse-coded as needed.

After completing all five blocks of ratings, participants were asked whether they had recognized any of the officials, and filled out a short survey questionnaire on demographic characteristics, political attitudes, and personality.

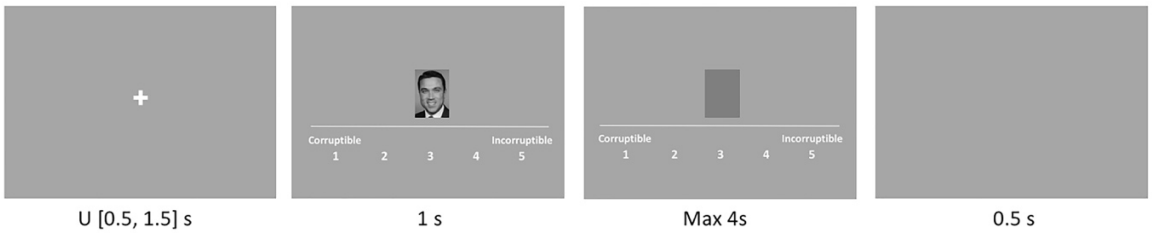


Fig. 1. An example trial in the corruptibility-judgment block. Each trial began with a fixation cross. Then a photo of an official appeared for 1 s. The orientation of the scale was randomly flipped for each block and each participant. Participants made a decision by pressing one of the number keys from "1" to "5" on their keyboard. As soon as a valid key was pressed (or 4 s after the photo disappeared if no valid key was pressed), the trial ended, and there was a blank inter-stimulus interval.

Results

Reliability of face-based trait inferences. Following our preregistered plan, we excluded from further analysis any responses faster than 100 ms and data for officials who were recognized. Among the 82 participants, 7 recognized one official (in total, four officials were recognized). The percentages of participants who used the full scale to rate the faces ranged from 59% to 68% across the five traits, and over 90% of the participants used scores on both sides of the midpoint to rate the faces (see Fig. S1 in the Supplemental Material available online).

First, we checked whether participants gave consistent judgments for a face across different traits. We expected consistent ratings for a face on traits with the same valence to

be positively correlated and ratings on traits with opposite valences to be negatively correlated. Although this was not planned in our preregistration, we computed repeated measures correlations (the R function rmcorr) to determine the common within- individuals correlations for ratings between each pair of traits, to handle the issue of nonindependence in repeated measures. Results (see Table S1 in the Supple- mental Material for coefficients and 95% confidence intervals, or CIs) showed that at an individual level, judgments of a face for traits with the same valence were positively correlated (repeated measures rs ranging from .24 to .31, ps < .001), and judgments of a face for traits with opposite valences were negatively correlated (repeated measures rs ranging from –.30 to –.21, ps < .001). Following our preregistered plan, we also analyzed the consistency of these ratings at an aggregate level. Ratings for each face were first averaged over participants, and (tie-corrected) Spearman correlation coefficients were calculated for each pair of traits with those averaged ratings. Aggregate-level judgments for a face were highly consistent across traits because they averaged out the measurement noise inherent in the individual-level correlations ($|r| \geq .75$; see Table S2 in the Supplemental Material).

Next, intraclass correlation coefficients (ICCs) were computed for each trait separately to test whether inferences of a trait showed consensus across participants— ICCs were computed according to type ICC(2, k) on the basis of complete cases. A high ICC indicates that the total variance in ratings is mainly explained by rating variance across images instead of across participants. In line with prior literature (see the introduction), our results showed high consensus among participants for inferences of corruptibility, ICC = .81, F(48, 3888) = 6.4, 95% CI = [.73, .88]; dishonesty, ICC = .82, F(45, 3645) = 6.7, 95% CI = [.74, .89]; selfishness, ICC = .86, F(42, 3402) = 8.1, 95% CI = [.80, .91]; trustworthiness, ICC = .82,

F(43, 3483) = 6.7, 95% CI = [.74, .89]; and generosity, ICC = .82, F(43, 3483) = 6.6, 95% CI = [.74, .89].

Association between corruption records and face-based trait inferences: preregistered analyses. Our primary interest in the current study was the extent to which trait inferences from a face were associated with actual corruption records. First, we followed the analysis methods planned in our preregistration and tested for these associations on the basis of inference judgments aggregated across participants and individually within subjects. For inferences of a negative trait, we deemed an official to be categorized accurately if he was convicted of corruption and received a high rating (> 3) or, conversely, if he had a clean record and received a low rating (≤ 3); for inferences of a positive trait, we deemed an official to be categorized accurately if he was convicted of corruption and received a low rating (< 3) or, conversely, if he had a clean record and received a high rating (≥ 3).

One-sample, one-tailed proportion tests against chance (50%) were performed on the aggregated-level accuracies across officials. One-sample one-tailed t tests against chance (50%) were performed on the individual-level accuracies across participants (we also calculated individual-level accuracies by categorizing midpoint 3 in the opposite way; see Table S3 in the Supplemental Material). Results (summarized in Table 1) showed that both aggregate-level and individual-level inferences of traits were associated with actual corruption records of the facial identities at a level better than chance (see Fig. S2 in the Supplemental Material for full distributions of individual-level accuracies).

Table 1. Results for Correctly Categorized Officials Based on Aggregate-Level Trait Inferences and Individual-Level Trait Inferences From Study 1

Trait	Aggregate-level accuracy: Percentage of correctly categorized officials ($N = 72$)	Aggregate-level accuracy: Lower bound of 95% CI	Aggregate-level accuracy: $\chi^2(1)$	Aggregate-level accuracy: p	Average individual-level accuracy[a]: Mean accuracy ($N = 82$)	Average individual-level accuracy[a]: SD	Average individual-level accuracy[a]: Lower bound of 95% CI	Average individual-level accuracy[a]: $t(81)$	Average individual-level accuracy[a]: Cohen's d
Corruptibility	69.44%	59.22%	10.13	< .001	55.73%	6.95%	54.46%	7.47	0.82
Dishonesty	70.83%	60.67%	11.68	< .001	54.82%	6.41%	53.64%	6.81	0.75
Selfishness	66.67%	56.36%	7.35	.003	55.10%	6.76%	53.86%	6.83	0.75
Trustworthiness	68.06%	57.79%	8.68	.002	55.03%	6.41%	53.85%	7.10	0.78
Generosity	63.89%	53.53%	5.01	.013	54.97%	5.99%	53.87%	7.51	0.83

Note: CI = confidence interval. [a]All *p*s for this variable are less than .001.

Association between corruption records and face-based trait inferences: extensions to preregistered analyses. Beyond our planned preregistered analyses, we conducted three additional robustness checks on the association between trait inferences from faces and corruption records. First, we confirmed that the above-chance accuracy we observed was not driven just by a small subset of faces: For each trait, we ranked the officials by the number of participants who categorized them accurately; we then calculated the average individual-level accuracy for subsets of stimuli in which the officials were progressively excluded one by one from the official who was accurately categorized by most participants to the official who was accurately categorized by fewest participants. For all five traits, average individual-level accuracies decreased smoothly as the highest ranked officials were removed and stayed above chance even after the 12th highest ranked official was excluded from the stimulus set (see Table S4 in the Supplemental Material).

Second, although participants were not informed of the purpose of the study or the percentage of corrupt politicians in our stimulus set (they were told only that these people were politicians), their beliefs (implicit or explicit) about the base rates of corrupt politicians in the real world or the percentage of corrupt politicians in our experiment might bias the ratings they gave. We corrected for such possibly idiosyncratic biases among our

participants by calculating individual-level accuracies using an alternative method. Ratings for each participant were centered on that participant's mean across all of his or her ratings on a trait (see Fig. S3 in the Supplemental Material for the full distributions of mean ratings).

For this analysis, inferences of a negative trait were deemed accurate if the official had been convicted of corruption and received a rating from a participant that was higher than the participant's mean rating or, conversely, if the official had a clean record and received a rating from a participant that was lower than the participant's mean rating; inferences of a positive trait were deemed accurate if the official was convicted of corruption and received a rating from a participant that was lower than the participant's mean rating or, conversely, if the official had a clean record and received a rating from a participant that was higher than the participant's mean rating. One-sample, one-tailed t tests against chance (50%) were performed on individual- mean-centered accuracies across participants. Corroborating the results reported previously, individual-level trait inferences correlated with officials' corruption records at a level better than chance, and the effect sizes were large—corruptibility inferences: $M = 55.57\%$, $SD = 7.75\%$, lower bound of 95% CI = 54.14%, $t(81) = 6.50$, $p < .001$, $d = 0.72$; dishonesty inferences: $M = 55.12\%$, $SD = 6.43\%$, lower bound of 95% CI = 53.94%, $t(81) = 7.22$, $p < .001$, $d = 0.80$; selfishness inferences: $M = 54.95\%$, $SD = 7.87\%$, lower bound of 95% CI = 53.50%, $t(81) = 5.69$, $p < .001$, $d = 0.63$; trustworthiness inferences: $M = 55.59\%$, $SD = 6.53\%$, lower bound of 95% CI = 54.39%, $t(81) = 7.75$, $p < .001$, $d = 0.86$; and generosity inferences: $M = 55.31\%$, $SD = 6.95\%$, lower bound of 95% CI = 54.03%, $t(81) = 6.92$, $p < .001$, $d = 0.76$.

Third, to address the concern that dichotomizing ratings into accurate and inaccurate might lead to loss of measurement sensitivity and to handle the nonindependence in ratings

due to repeated measures designs, we performed general linear mixed-model (GLMM) analyses for inferences of each trait, respectively. Officials' corruption records (1 = conviction, 0 = clean) were regressed on individual-level ratings in logistic models, and participants were treated as random factors (N = 5,757; N was determined by the number of participants multi- plied by the number of faces, excluding omitted observations; observations from a participant for a face would be omitted if ratings were not available for all five traits). In addition, photo characteristics (the official's age and smile intensity; the presence of glasses, a beard, a mustache, and a bald head; image clarity; and image sources) were included as control variables in all models. All continuous variables were standardized.

We observed significant effects of trait ratings: Officials who were rated as looking more corruptible, $b = 0.23$, $SE = 0.03$, 95% CI = [0.17, 0.29], $z = 7.66$, $p < .001$; dis- honest, $b = 0.17$, $SE = 0.03$, 95% CI = [0.11, 0.23], $z = 5.75$, $p < .001$; and selfish, $b = 0.20$, $SE = 0.03$, 95% CI = [0.14, 0.26], $z = 6.77$, $p < .001$, were more likely to have been convicted of corruption, whereas officials who were rated as looking more trustworthy, $b = -0.19$, $SE = 0.03$, 95% CI = [−0.25, −0.13], $z = -6.41$, $p < .001$, and generous, $b = -0.20$, $SE = 0.03$, 95% CI = [−0.26, −0.14], $z = -6.59$, $p < .001$, were less likely to have been convicted of corruption (for complete lists of coefficients, see Table S5 in the Supplemental Material).

Association between corruption records and face-based trait inferences: further exploration of potential mechanisms. Finally, we performed two additional analyses that were also beyond our preregistration. We performed GLMM analyses on two subsets of data to test two photo-selection–related mechanisms underlying the face–corruption-record association we found. To test the hypothesis that potential negative biases in the convicted officials' photos that were from sources beyond the control of the officials might be driving

the association, we conducted GLMM analyses on a subset of data that included only officials whose photos were self-selected—that is, those from government websites and personal campaign websites (n = 45; 20 were convicted of corruption; in this subset, only 1 official had a beard, and only 2 officials were bald, and therefore these two predictors were removed from the model).

The associations between trait inferences and records remained significant— corruptibility inferences: b = 0.24, SE = 0.04, 95% CI = [0.17, 0.32], z = 6.81, p < .001; dishonesty inferences: b = 0.19, SE = 0.04, 95% CI = [0.12, 0.26], z = 5.21, p < .001; selfishness inferences: b = 0.18, SE = 0.04, 95% CI = [0.11,0.25], z = 5.07, p < .001; trustworthiness inferences: b = −0.20, SE = 0.04, 95% CI = [–0.27, –0.13], z = −5.63, p < .001; and generosity inferences: b = −0.17, SE = 0.04, 95% CI = [–0.24, –0.10], z = −4.66, p < .001.

To test the hypothesis that potential negative biases in the convicted officials' photos that were taken after conviction might be driving the face–corruption-record association, we conducted GLMM analyses on a subset of data that included only officials whose photo dates were known (and were prior to the date of conviction, for convicted officials; n = 62; 26 were convicted of corruption). The associations between trait inferences and records became weaker but remained significant— corruptibility inferences: b = 0.17, SE = 0.03, 95% CI = [0.10, 0.23], z = 4.93, p < .001; dishonesty inferences: b=0.11,SE=0.03,95%CI=[0.04,0.18],z=3.29,p= .001; selfishness inferences: b = 0.16, SE = 0.03, 95% CI = [0.09, 0.22], z = 4.64, p < .001; trustworthiness inferences: b = −0.14, SE = 0.03, 95% CI = [–0.20, –0.07], z = −4.06, p < .001; and generosity inferences: b = −0.19, SE = 0.03, 95% CI = [–0.26, –0.13], z = −5.74, p < .001. This indicates that while potential

biases in photo selection can explain some of the relationship between trait ratings and officials' records, they cannot entirely account for our main findings.

Two additional analyses were preregistered but are not presented in this article; the codes to conduct those analyses can be found at https://osf.io/k4mds/. In our preregistration, we proposed an alternative approach to analyze individual-level ratings (logistic regression with adjusting standard errors for clustering). These analyses are not presented here because the GLMM analyses reported previously are more appropriate for handling repeated measures. We had also planned analyses of correlations between individual-level accuracies and response times, but these were intended to answer a question that is beyond the scope of the current article.

3.3 Study 2

Study 1 showed that compared with peers with clean records, federal and state officials who were convicted of political corruption were perceived as more corruptible, dishonest, and selfish and less trustworthy and generous. To assess the generalizability of these findings, we next tested whether they would also hold for officials from lower levels of governments, and for the comparison between officials with clean records and officials who violated campaign finance laws.

Method

Participants. This study was preregistered before data collection began (https://osf.io/tgzpz/). A pilot study with 24 MTurk workers conducted in February 2017 yielded an estimated effect size of 1.39, justifying a minimum sample size of 10 participants.

To ensure sufficient power and to have a sample size comparable with that of Study 1, we predetermined the sample size to be 100 participants. The same inclusion and exclusion criteria as in Study 1 were applied (including exclusion of participants from Study 1). We excluded 22 individuals, 3 for not being native English speakers, 2 for pressing the same response key for all trials in a block, and 17 for failing to input valid responses for more than 10% of the trials in a block. After these exclusions, there were 78 MTurk workers who participated in this study in February and March 2017 (33 female; age: M = 38 years, SD = 11; 83% White, 9% Black, 6% Asian).

Stimuli. Stimuli were photos of 80 real elected officials. All officials were Caucasian males who held offices in California state and local governments. Photos were official headshots obtained from government websites and personal campaign websites (86%), news articles, and Wikipedia (14%). All photos were converted to grayscale on a plain gray background and were cropped to a uniform size. All faces were frontal, smiling, in clear focus, and centered in the middle of the image.

Among the 80 officials, half violated the California Political Reform Act (officials with violations), and the other half had clean records (officials without violations). The officials with violations were from the data- base of the California Fair Political Practices Commission's "Enforcement Cases" (http://www.fppc.ca.gov/aboutfppc/hearings-meetings-workshops/current-agenda/ past-agendas.html). To reduce sources of variability, we included only officials who were Caucasian, were male, and had committed a violation related to election campaigns (laundered campaign contributions, accepted over-the-limit gifts and contributions, improperly used campaign funds, had conflicts of interest, inadequately or inaccurately reported on campaign statements, did not file campaign

statements or filed them late, or were involved in illegal campaign coordination). In addition, we included only successful candidates of the election related to the violation, whose cases merited pursuit of a fine over $215, whose cases were closed between January 2015 and January 2017, whose age information was publicly available, and who had publicly available frontal photographs of acceptable clarity that featured them smiling. All photographs had been taken while in office. Most photos of the officials with violations had a known creation date, and we confirmed that the photos were taken before the cases were closed (88%); for the rest of the photos (12%), the creation date was unknown (analyses were also performed when excluding data for these stimuli). The officials without violations were randomly generated from our available stimulus set from the list of incumbents who had clean records and were holding the same office in the state of California and were the same gender, the same race, and of similar age as the officials with violations.

Procedure. Participants followed the same experimental procedure as in Study 1 but viewed a new set of stimuli, as described previously.

Results

Reliability of face-based trait inferences. Following our preregistered plan, we excluded responses faster than 100 ms and responses for officials who were recognized. Among the 78 participants, only 1 recognized one official. As in Study 1, ratings across faces given by each participant had sufficient variance: The majority of participants used the full scale to rate the faces (the percentages of participants ranged from 58% to 63% across the five traits), and more than 97% of the participants used scores on both sides of the midpoint to rate the faces (see Fig. S4 in the Supplemental Material).

To test how consistently a participant judged a face across different traits, we computed repeated measures correlations (using the R function rmcorr) following the method in Study 1. A participant's ratings of a face on traits with the same valence were positively correlated (repeated measures rs ranging from .26 to .35, ps < .001), and ratings on traits with opposite valences were negatively correlated (repeated measures rs ranging from –.38 to –.26, ps < .001; see Table S6 in the Supplemental Material for coefficients and 95% CIs). As planned in our preregistration, we also computed (tie-corrected) Spearman correlation coefficients for each pair of traits using ratings averaged over participants for each face. Aggregate-level judgments of a face were once again highly consistent across traits ($|r| \geq .77$; see Table S7 in the Supplemental Material).

In line with Study 1 and prior literature, we observed high consensus among participants for face-based judgments of corruptibility, ICC = .81, F(64, 4928) = 6.3, 95% CI = [.75, .87]; dishonesty, ICC = .82, F(61, 4697) = 6.7, 95% CI = [.75, .87]); selfishness, ICC = .82, F(45, 3465) = 6.2, 95% CI = [.74, .89]; trustworthiness, ICC = .86, F(58, 4466) = 8.6, 95% CI = [.81, .91]; and generosity, ICC = .87, F(56, 4312) = 8.9, 95% CI = [.82, .91]. ICCs were computed according to type ICC(2, k) on the basis of complete cases.

Association between records of violations and face-based trait inferences: preregistered analyses. Following the methods in Study 1, we calculated the proportions of correctly categorized officials for each trait on the basis of aggregate-level inferences and individual-level inferences as planned in our preregistration. Table 2 summarizes one-sample one-tailed proportion-test statistics of aggregate-level accuracies and one-sample one-tailed t-test statistics of individual-level accuracies (see Fig. S5 in the Supplemental Material for full distributions of individual- level accuracies; see Table S8 in the Supplemental Material

for average individual-level accuracies calculated with categorizing midpoint 3 in an opposite way). The findings replicated those from Study 1.

Table 2. Results for Correctly Categorized Officials Based on Aggregate-Level Trait Inferences and Individual-Level Trait Inferences From Study 2

	Aggregate-level accuracy				Average individual-level accuracy[a]				
Trait	Percentage of correctly categorized officials ($N = 80$)	Lower bound of 95% CI	$\chi^2(1)$	p	Mean accuracy ($N = 78$)	SD	Lower bound of 95% CI	$t(77)$	Cohen's d
Corruptibility	67.50%	57.79%	9.11	.001	54.72%	6.59%	53.48%	6.32	0.72
Dishonesty	70.00%	60.38%	12.01	< .001	56.15%	6.51%	54.92%	8.34	0.94
Selfishness	65.00%	55.23%	6.61	.005	55.78%	7.21%	54.42%	7.08	0.80
Trustworthiness	70.00%	60.38%	12.01	< .001	56.00%	6.31%	54.74%	7.98	0.90
Generosity	67.50%	57.79%	9.11	.001	55.80%	5.51%	54.76%	9.29	1.05

Note: CI = confidence interval. [a]All *p*s for this variable are less than .001.

Association between corruption records and face-based trait inferences: extensions to preregistered analyses. As in Study 1, we conducted three analyses in addition to those we had preregistered to check the robustness of the association between trait inferences from officials' faces and the records of violations of the facial identities. First, we verified that the above-chance accuracy observed earlier was not driven just by a small subset of faces. Following the same approach as Study 1, we recalculated individual-level accuracies for subsets of stimuli in which the stimulus was excluded one by one from the official who was accurately categorized by most participants to the official who was accurately categorized by the fewest participants. Average individual-level accuracies for each trait decreased smoothly as the highest ranked officials were progressively excluded and stayed above chance even after the 14th highest ranked official was excluded from the stimulus set (see Table S9 in the Supplemental Material).

Second, participants' beliefs (implicit or explicit) about the base rates of corrupt politicians in the real world or the percentage of corrupt politicians in our study might have influenced their trait judgments from politicians' faces. Consequently, we computed the individual-level accuracies using an alternative method that took into account the heterogeneous beliefs of base rates across participants. As in Study 1, a mean rating was computed for each participant by averaging the ratings he or she gave across all faces for a trait (see Fig. S6 in the Supplemental Material for the full distributions of mean ratings). This mean rating was used as a cutoff for dichotomizing whether the participant's rating correctly categorized an official. These individual-mean-centered accuracies across participants were then tested against chance (50%). We observed significantly above-chance accuracies and large effect sizes for corruptibility inferences, $M = 55.06\%$, $SD = 6.98\%$, lower 95% CI = 53.74%, $t(77) = 6.40$, $p < .001$, $d = 0.72$; dishonesty inferences, $M = 56.06\%$, $SD = 7.32\%$, lower 95% CI = 54.68%, $t(77) = 7.31$, $p < .001$, $d = 0.83$; selfishness inferences, $M = 55.74\%$, $SD = 7.98\%$, lower 95% CI = 54.24%, $t(77) = 6.36$, $p < .001$, $d = 0.72$; trust- worthiness inferences, $M = 56.00\%$, $SD = 7.05\%$, lower 95% CI = 54.67%, $t(77) = 7.52$, $p < .001$, $d = 0.85$; and generosity inferences, $M = 55.61\%$, $SD = 6.62\%$, lower 95% CI = 54.36%, $t(77) = 7.48$, $p < .001$, $d = 0.85$.

Third, data were further analyzed in GLMM analyses to handle the nonindependence in ratings due to the repeated measures design and avoid any data dichotomization. Officials' records of violations (1 = violation, 0 = clean) were regressed on individual-level ratings in logistic models, and participants were treated as random factors (N = 6,115; N was determined by the number of participants multiplied by the number of faces excluding omitted observations; observations from a participant for a face would be omitted if ratings

were not available for all five traits). In addition, photo characteristics (the official's age and smile intensity; the presence of glasses, a beard, a mustache, and a bald head; image clarity; and image sources) were included as control variables in all models. All continuous variables were standardized. Results revealed significant effects of trait ratings: Officials who were rated as looking more corruptible, b = 0.24, SE = 0.03, 95% CI = [0.18, 0.29], z = 8.19, p < .001; dishonest, b = 0.28, SE = 0.03, 95% CI = [0.23, 0.34], z = 9.77, p < .001; and selfish, b=0.27,SE=0.03,95%CI=[0.21,0.32],z=9.31,p< .001, were more likely to have violated campaign finance laws, whereas officials who were rated as look- ing more trustworthy, b = −0.26, SE = 0.03, 95% CI = [–0.32, –0.20], z = −9.05, p < .001, and generous, b = −0.27, SE = 0.03, 95% CI = [–0.33, –0.22], z = −9.53, p < .001, were less likely to have violated campaign finance laws (for complete lists of coefficients, see Table S10 in the Supplemental Material).

Association between corruption records and face-based trait inferences: further exploration of potential mechanisms. Finally, to elucidate whether the observed associations between trait judgments from faces and records of violations of the facial identities might in part be attributable to unintended properties of photo sources, we performed GLMM analyses on two subsets of data, respectively. For one subset of data, we excluded officials whose photos were not self-selected—that is, we included only officials whose photos were from government websites and personal campaign websites (N = 69; 33 violated campaign finance laws). Trait inferences based on photos self-selected by the officials were significantly associated with the officials' records of violations—corruptibility inferences: b = 0.23, SE = 0.03, 95% CI = [0.17, 0.29], z = 7.48, p < .001; dishonesty inferences: b = 0.26, SE = 0.03, 95% CI = [0.20, 0.32], z = 8.68, p < .001; selfish-

ness inferences: b = 0.25, SE = 0.03, 95% CI = [0.19, 0.31], z = 8.37, p < .001; trustworthiness inferences: b = −0.25, SE = 0.03, 95% CI = [–0.31, –0.19], z = −8.31, p < .001; and generosity inferences: b = −0.25, SE = 0.03, 95% CI = [–0.31, –0.19], z = −8.15, p < .001.

To test the hypothesis that potential negative biases in photos of officials with violations if the photos were taken after the violation was caught might be driving the face–corruption-record association, we performed GLMM analyses on a subset of data that included only officials for whom the dates on which their photo was taken was known (and were taken prior to the date when the violation was caught, for officials with violations; n = 75; 35 violated campaign finance laws). The associations between trait inferences and records remained significant: corruptibility inferences, b = 0.25, SE = 0.03, 95% CI = [0.19, 0.31], z = 8.33, p < .001; dis- honesty inferences, b = 0.29, SE = 0.03, 95% CI = [0.24, 0.35], z = 9.78, p < .001; selfishness inferences, b = 0.29, SE = 0.03, 95% CI = [0.23, 0.34], z = 9.56, p < .001; trustworthiness inferences, b = −0.28, SE = 0.03, 95% CI = [–0.33, –0.22], z = −9.27, p < .001; and generosity inferences, b = −0.30, SE = 0.03, 95% CI = [–0.36, –0.24], z = −10.17, p < .001.

The analysis of the correlation between individual- level accuracies and response times was also planned. Results are not detailed here because these analyses intended to answer a question that is beyond the scope of the current article. For readers interested in these results, all relevant data and analysis codes can be accessed at https://osf.io/k4mds/.

3.4 Study 3

Study 2 replicated the face–record association found in Study 1 with an independent set of stimuli. However, these findings were based on traits that either were close in meaning to corruptibility (selfishness, dishonesty) or have the opposite meaning from it (trustworthiness, generosity). This resulted in our findings deriving from a single underlying factor with no comparison to different traits. Study 3 therefore aimed to test that the effects found in Studies 1 and 2 could be attributed specifically to corruptibility judgments.

Method

Participants. This study was preregistered before data collection began (https://osf.io/7a7eu/). To ensure a sample size comparable with that used in Study 1, we recruited 100 participants via MTurk. The same inclusion and exclusion criteria as in Study 1 were applied; in addition, participants were required to have no prior participation in Study 1. We excluded 15 individuals, 2 for not being native English speakers, 2 for pressing the same response key for all trials in a block, and 11 for failing to input valid responses for more than 10% of the trials in a block. After exclusions, data were retained from 85 participants who were recruited from MTurk in February and March 2017 (42 female; age: M = 37 years, SD = 10; 88% White, 6% Black, 4% Asian).

Stimuli and procedure. We used stimuli identical to those from Study 1 and a protocol similar to that of Study 1 except that participants evaluated the officials on a different set of traits: corruptibility, aggressiveness, masculinity, competence, and ambitiousness.

Results

Reliability of face-based trait inferences. We excluded from further analysis any responses faster than 100 ms and responses for officials who were recognized. Among the 85 participants, 3 recognized at least one official (in total, two officials were ever recognized).

We first checked the variation of individual-level ratings across faces for each trait. For all the four traits except masculinity, the majority of participants used the full scale to rate the faces, and over 92% of the participants used scores on both sides of the midpoint to rate the faces (see Fig. S7 in the Supplemental Material). Not surprisingly, given that all the officials were male, ratings for masculinity were skewed toward masculine; however, 80% of the participants still rated the faces on masculinity using scores on both sides of the midpoint.

Participants showed high consensus on face-based trait judgments for corruptibility, ICC = .86, $F(52, 4368) = 8.7$, 95% CI = [.80, .91]; aggressiveness, ICC = .85, $F(54, 4536) = 8.3$, 95% CI = [.79, .90]; masculinity, ICC = .89, $F(53, 4452) = 13.6$, 95% CI = [.85, .93]; and competence, ICC = .84, $F(58, 4872) = 8.4$, 95% CI = [.78, .89]; and the consensus on ambitiousness judgments was fair, ICC = .69, $F(53, 4452) = 3.9$, 95% CI = [.57, .79]. ICCs were computed according to type ICC(2, k) based on complete cases.

Association between corruption records and face-based trait inferences. Critically, we replicated the results found in Study 1 with this new set of participants: Officials who were convicted of political corruption looked more corruptible than their peers with clean records, aggregate- level accuracy = 72.22%, lower 95% CI = 62.12%, $\chi2(1) = 13.35$, $p < .001$; average individual-level accuracy = 56.30%, SD = 7.22%, lower 95% CI = 55.00%, $t(84) = 8.04$, $d = 0.87$, $p < .001$. Additionally, participants in Study 1 and the present study viewed the same set of stimuli, and their judgments (averaged over participants within each study) of how corruptible a face looked were highly correlated, $\rho = 0.88$, 95% CI = [0.81, 0.92], $p < .001$.

Interestingly, data from the present study revealed that officials who were perceived as more aggressive were also more likely to have been convicted of political corruption, aggregate-level accuracy = 66.67%, lower 95% CI = 56.36%, $\chi 2(1) = 7.35$, $p = .003$; average individual- level accuracy = 55.09%, SD = 6.13%, lower 95% CI = 53.98%, $t(84) = 7.66$, $d = 0.83$, $p < .001$. However, the associations between corruption records and inferences of masculinity, ambitiousness, and competence were not statistically reliable (95% CIs included 50%, and ps were > .01 for aggregate-level accuracies).

Correlation structure of trait inferences. Our primary interest in the current study was whether the observed face–corruption-record associations resulted from inferences of specific traits or global valence evaluations of the face. We first analyzed the correlation structure of the trait inferences. To allow for analyses across all nine traits (those from Study 1 and Study 3 combined), we first averaged inferences of traits across participants for each face, and then these aggregate-level data were merged across Study 1 and the present study. Figure 2 shows the Spearman correlation coefficients between each pair of traits. All correlations were in expected directions and generally strong, except for masculinity and ambitiousness.

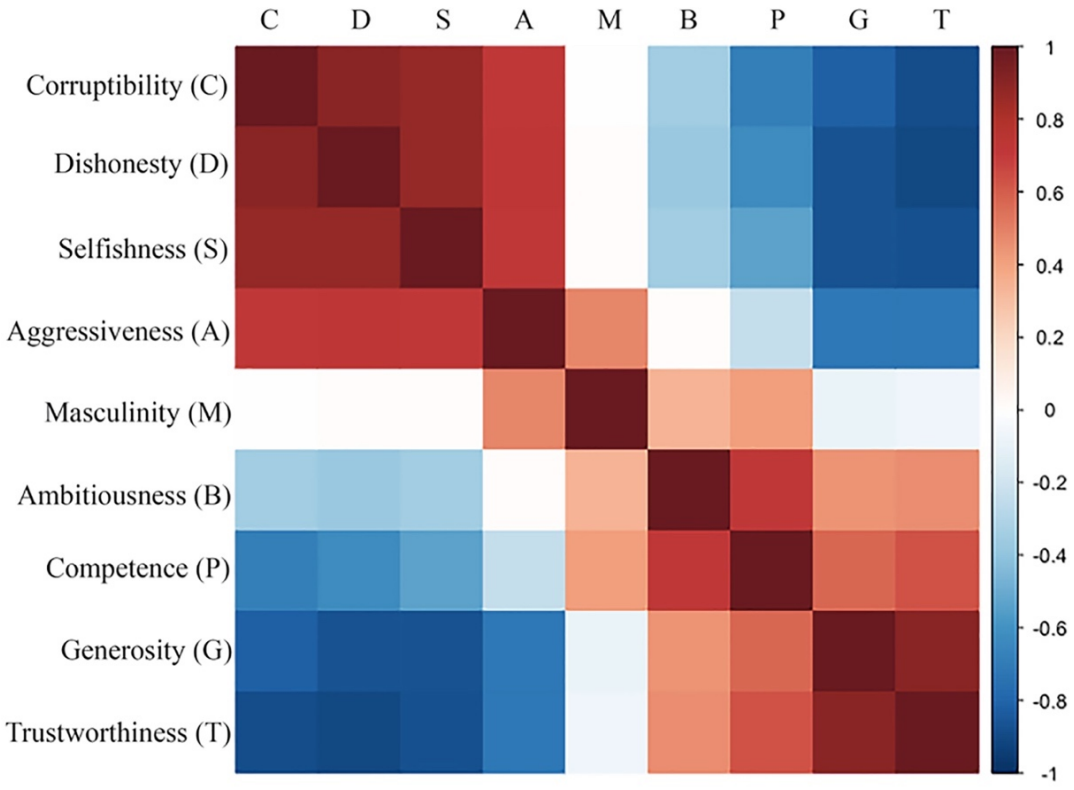


Fig. 2. Spearman correlation coefficients between each pair of traits across Study 1 and Study 3, calculated with aggregate-level trait ratings ($N = 72$). Inferences of corruptibility were averaged over the two studies.

A principal component analysis with varimax rotation indicated that these trait inferences clustered on three distinctive factors: a corruptibility-related factor (corruptibility, dishonesty, selfishness, aggressiveness, generosity, and trustworthiness), a competence-related factor (competence and ambitiousness), and a masculinity-related factor (masculinity), each accounting for 57%, 19%, and 15% of the variance in the data, respectively (see Table S11 in the Supplemental Material). A composite score was computed for each factor with the trait inferences that comprised it (Todorov et al., 2005; for the corruptibility-related factor, positive and negative traits were aggregated with

opposite signs). Importantly, logistic regression analyses with each of these three factors independently while controlling for other covariates demonstrated that only the corruptibility-related factor was associated with corruption records (Fig. 3).

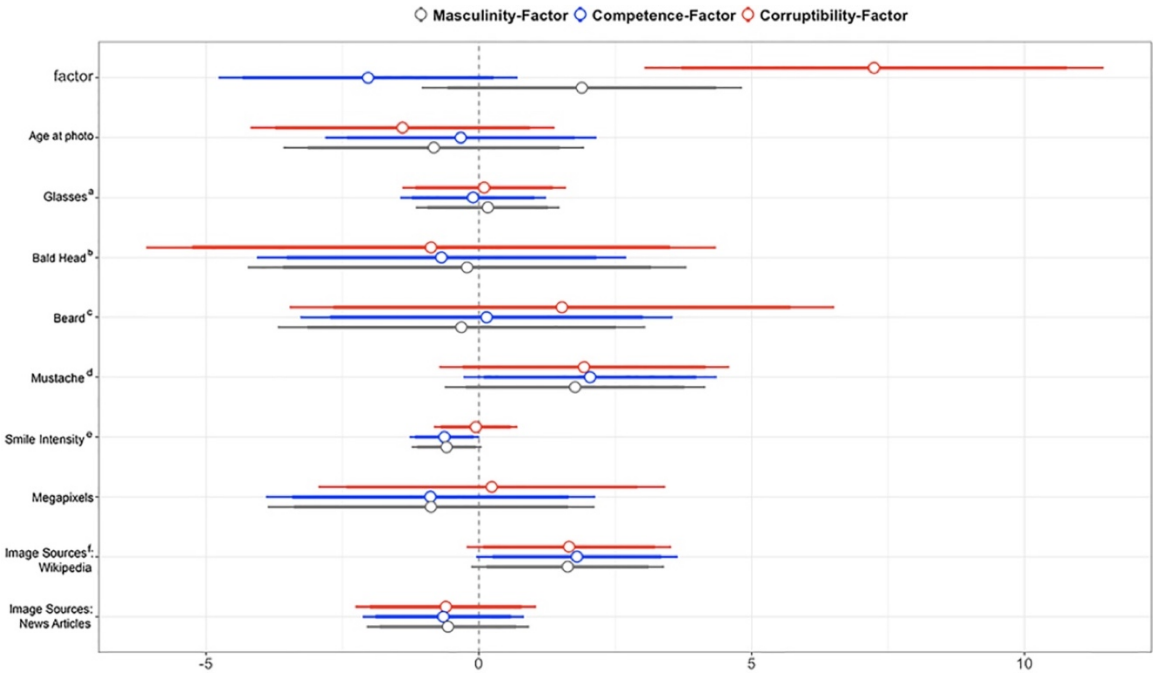


Fig. 3. Unstandardized logistic regression (N = 72) coefficients for factors and photo characteristics as regressors of the officials' corruption records (1 = conviction, 0 = clean) in Study 3. Thick lines represent 90% confidence intervals, and thin lines represent 95% confidence intervals. Glasses is a dummy variable with 1 indicating the official wore glasses. Bald head is a dummy variable with 1 indicating the official was bald. Beard is a dummy variable with 1 indicating the official had a beard. Mustache is a dummy variable with 1 indicating the official had a mustache. Smile intensity was coded manually with three levels (1 = smile with no teeth exposed, 2 = smile with teeth but not gums exposed, 3 = smile with gums exposed). There were three sources of photos: government and campaign websites (benchmark), Wikipedia, and news articles. All variables were normalized into the range of [0, 1].

3.5 Study 4

Study 3 demonstrated that elected officials' corruption records were associated with specific trait inferences (e.g., corruptibility). A final important question concerns the facial features that make some officials look more corruptible than others. Study 4 provided a preliminary exploration of this question by first estimating the relationship between objective facial structures, inferences of traits, and officials' records with causal mediation models (Study 4a; not preregistered). Second, the causal effects suggested by the mediation analysis were directly tested in an experiment that manipulated the face stimuli (Study 4b; preregistered).

Study 4a

Method. Study 1 and Study 2 collected judgments of a common set of traits (corruptibility, dishonesty, selfish- ness, trustworthiness, and generosity) for two distinct sets of officials. The present study merged data from both studies. For trait judgments of an official given by a participant, we computed a composite score using his ratings across the five traits and referred to it as corruptibility- related trait inferences.

Officials were those used in Study 1 and Study 2. Whether an official is corrupt was measured by his record. A record of conviction of political corruption or violation of campaign finance laws suggests that an official is corrupt, and a clean record suggests an official is not corrupt. Officials' records are one metric of real-world corruption, but the potential measurement error of this metric is beyond the scope of the present study.

Eight metrics representing the distances between facial landmarks specified by anthropometric definitions were measured (Farkas, 1994; Stirrat & Perrett, 2010). Stimuli were the photos of the elected officials used in Study 1 and Study 2. We adjusted for shifts

in posture or tilts in head angle by making all measurements only on one side of the face—the side turned most toward the camera—and by generating a face-based reference frame—the horizontal axis of the face was defined by the line connecting the two pupils, and the vertical axis was defined by the line through landmark n (for nasion) that was perpendicular to the horizontal axis (Fig. 4). Summary statistics of these metrics are reported in the Supplemental Material (Table S12).

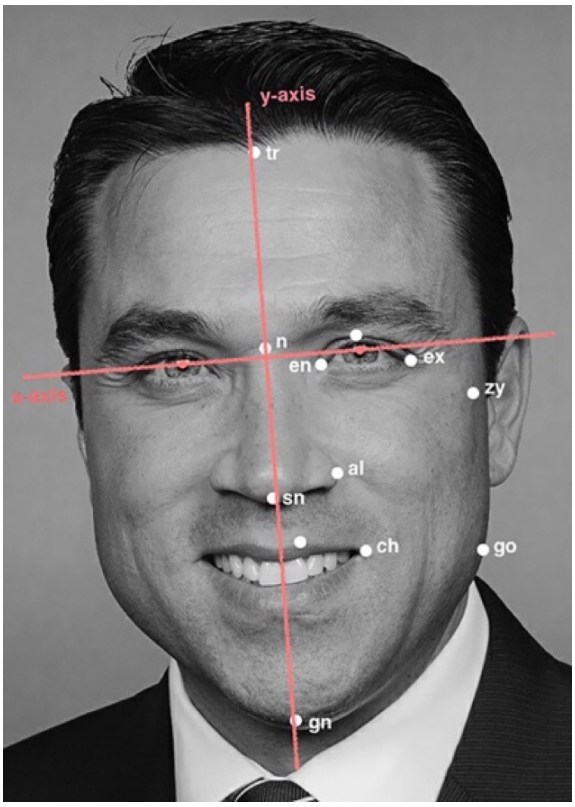


Fig. 4. Illustration of facial landmarks (white points) and the coordinate system (red lines). Facial width-to-height ratio was calculated as the bizygomatic width (the horizontal distance from landmark *zy* to the *y*-axis multiplied by 2) divided by the upper-face height (the vertical distance from the highest point of the upper lip to the highest point of the eyelids). Face width/lower-face height was calculated as the bizygomatic width divided by the lower-face height (the vertical distance between landmark *ex* and landmark *gn*).

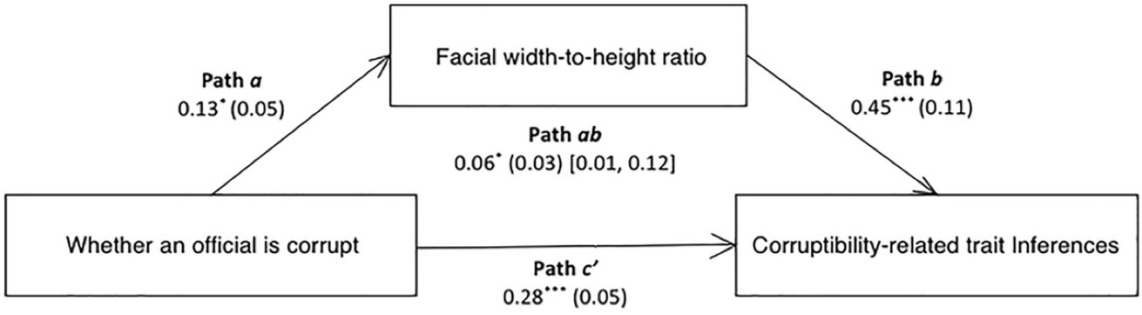
Lower face/face height was calculated as the lower-face height divided by the physiognomic face height (the vertical distance between landmark *tr* and landmark *gn*). Cheekbone prominence was calculated as the bizygomatic width divided by the jawbone width (the horizontal distance from landmark *go* to the *y*-axis multiplied by 2). Internal eye-corner distance was calculated as the ratio of the internal eye-corner width (the horizontal distance from landmark *en* to the *y*-axis multiplied by 2) to the bizygomatic width. Nose height was calculated as the ratio of the nose length (the vertical distance from landmark *n* to landmark *sn*) to the lower face height. Mouth width was calculated as the ratio of the mouth corner distance (the horizontal distance from landmark *ch* to the *y*-axis multiplied by 2) to the jawbone width. Nose/mouth width was calculated as the ratio of the nose width (the horizontal distance from landmark *al* to the *y*-axis multiplied by 2) to the mouth corner distance.

Results. Studies 1 to 3 demonstrated that officials who had clean records were judged differently on corruptibility than officials who were convicted of political corruption and those who violated campaign finance laws. To test the hypothesis that the perceptual difference was mediated by certain facial structures, we analyzed a causal mediation model linking whether an official is corrupt, corruptibility-related trait inferences, and each facial structure with data from Study 1 and Study 2 (Fig. 5). The effect of whether an official is corrupt on the facial structure (path a) and the effect of the facial structure on corruptibility-related trait inferences controlling for whether an official is corrupt (path b) constitute the indirect effect from whether an official is corrupt to corruptibility-related trait inferences (path ab). Path a was estimated with linear regression models. Path b was estimated with linear mixed models in which subjects, images nested within record types, and the interactions between subjects and record types were treated as random factors. The indirect effect was estimated with RMediation in R. The direct effect (path c′) of whether an official

is corrupt on corruptibility-related trait inferences after controlling for the indirect effect was estimated in the same model as for path b. Photo characteristics (the official's age and smile intensity; the presence of glasses, a beard, a mustache, and a bald head; image clarity; and image sources) were included as covariates in all models; for simplicity, these paths are not depicted in the figure.

Two of the eight facial structures were identified to have significant indirect effects: facial width-to-height ratio (unstandardized coefficient for path ab = 0.06, SE = 0.03, 95% CI = [0.01, 0.12]), and face width/lower face height (unstandardized coefficient for path ab = 0.11, SE = 0.04, 95% CI = [0.04, 0.18]). These results revealed that compared with officials who had clean records, those who were convicted of political corruption and violated campaign finance laws were perceived more negatively (more corruptible, dishonest, and selfish and less trustworthy and generous), and these negative impressions were partially attributable to higher facial width-to-height ratio and face width/lower-face height.

a



b

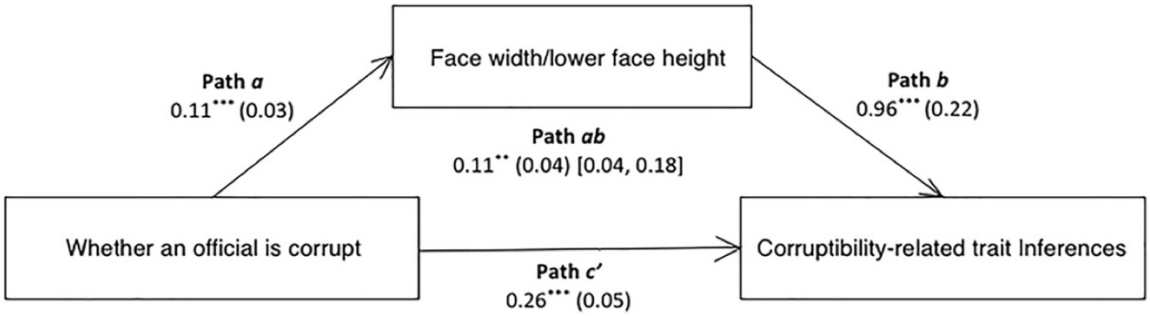


Fig. 5. Results of causal mediation analyses showing the influence of whether an official is corrupt on corruptibility-related trait inferences, as mediated by facial structures (Study 4a). A mediation model was constructed for each of the eight facial metrics separately and was tested with data from Study 1 and Study 2. Two of the eight facial metrics, (a) face width-to-height ratio and (b) face width/lower-face height, showed significant indirect effects. Unstandardized coefficients are shown, and standard errors are given in parentheses. Coefficients for path *a* were estimated in linear regression models. Coefficients for path *b* and path *c*′ were estimated in linear mixed models. The indirect effects of path *ab* were estimated with *RMediation* in R. Photo characteristics were included as covariates in all models; for simplicity, these variables and the corresponding paths are not depicted in the figure. No indirect effect was found for the other six facial metrics. Asterisks indicate significant paths (*$p < .05$, **$p < .005$, ***$p < .0005$). CI = confidence interval.

Study 4b

Study 4a suggests that compared with officials with slimmer faces, officials with wider faces were judged more negatively on corruptibility-related traits. This finding raises an important question: Given the same elected official, is how corruptible he looks influenced by how wide his face is in a photo? Study 4b directly tested this causal hypothesis by manipulating the facial width of the photos and contrasting the degree of corruptibility inferred from the slim, original, and fat version photos of the same official.

Method.

Stimuli. This study was preregistered before data collection began (https://osf.io/58x6e/). Stimuli were 450 black-and-white headshots of real elected officials. There were three versions of the stimuli: original, fat, and slim. Original stimuli consisted of 71 photos from Study 1 and 79 photos from Study 2 (1 photo from Study 1 and 1 photo from Study 2 were excluded from the present study because the manipulation of face width distorted these two faces). These 150 original stimuli were further manipulated with the Adobe Photoshop Face-Aware Liquify tool to increase face width by 7% and decrease face width by 7%, which resulted in two additional versions of each facial identity (see Fig. 6 for an example; all stimuli used in the present study can be accessed at https://osf.io/k4mds/). Fat stimuli consisted of the 150 photos with increased face width, and slim stimuli consisted of the 150 photos with decreased face width. This percentage of face-width change was the maximum manipulation we could achieve subject to the constraints that all faces should look natural and the manipulation should be subtle enough to go unnoticed.

a b c

Fig. 6. Example of the same face in (a) slim, (b) original, and (c) fat versions.

Participants. To investigate our main hypothesis that the same official would be judged as more corruptible when his face was fatter relative to when it was slimmer, we conducted a pilot study on MTurk with 18 participants, which yielded 2,700 observations. These observations gave an estimated effect size of 0.09, justifying a minimum sample size of 16 participants and 2,375 observations. To ensure sufficient power even with data exclusion, we predetermined the sample size to be 100 participants. Participants were required to be located in the United States, to be 18 years old or older, and to have normal or corrected-to-normal vision, an educational attainment of high school or above, a HIT approval rate greater than or equal to 95%, and no prior participation in the pilot study. Additionally, two open-ended questions and one closed-ended question in the survey at the end of the experiment (Table S13 in the Supplemental Material) gauged whether participants noticed that the width of the faces was manipulated; participants who mentioned face width to any of the open-ended questions were excluded from data analysis.

Only 1 participant recognized that the face width of the stimuli was manipulated; this individual was excluded from data analyses. Another 19 participants were excluded for failing to input valid responses for more than 10 trials. After exclusions, the final sample consisted of 80 participants (37 female; age: M = 38 years, SD = 10; 89% White, 5% Black, 5% Asian), who were recruited from MTurk in July 2017.

Procedures. Participants were not informed about the purpose of the study (they were told only that this was a study about judging how corruptible politicians looked on the basis of their photos). They were instructed to make their decisions as quickly and precisely as

possible. Participants viewed and evaluated the 450 stimuli one at a time in 10 blocks; they had the option to take breaks between blocks. The order of the 450 stimuli was randomized under a constraint that different versions of photos of the same facial identity did not appear within 10 consecutive images. Participants indicated how corruptible each face looked on a 9-point Likert scale anchored with bipolar adjectives (corruptible and incorruptible) and were encouraged to use the full range to rate the faces. The orientation of the scale was randomized for each participant. Except for the change of the Likert scale, the present study followed the same experimental procedures as the previous studies. After rating all stimuli, participants were asked whether they had recognized any of the officials or noticed that the width of the faces were manipulated (see Table S13) and filled out a short survey questionnaire on demographic characteristics, political attitudes, and personality.

Results. The survey responses revealed that when told that photos of the same politicians were shown more than once during the experiment and asked whether photos of the same politician were used, 61 of the 80 participants reported that they believed the repeated photos of the same politician were identical. The rest of the participants either indicated that they were not sure whether the photos of the same politician were different or identical (n = 3) or mentioned that the faces in these photos might have different facial expressions (n = 4), hair or facial hair (n = 4), smile intensity (n = 3), eyes (n = 1), glasses (n = 1), or head shapes (n = 1); the individual pictured might be wearing different clothing (n = 1); or some photos looked scarier (n = 1), might be mixed with parts from other pictures (n = 1), or might be taken from different angles (n = 1).

Most of the participants (71.25%) used the full range to rate the faces as instructed, and all participants used both sides of the scale to rate the faces. Data were analyzed in linear

mixed models, and subjects, images, and the interactions between subjects and versions were treated as random factors. As hypothesized, individual-level data showed that face width had a significant effect on inferences of corruptibility; specifically, a participant perceived an official as more corruptible when his face was fat relative to when his face was slim, $b = 0.06$, $SE = 0.02$, 95% CI = [0.03, 0.09], $p < .001$, $d = 0.22$. This unconscious perceptual bias was symmetrically driven by increasing face width, $b = 0.06$, $SE = 0.02$, 95% CI = [0.02, 0.11], $p = .008$, $d = 0.22$, and decreasing face width, $b = 0.06$, $SE = 0.03$, 95% CI = [0.02, 0.11], $p = .025$, $d = 0.22$. We further analyzed whether the perceptual bias to rate the fat version of a face as more corruptible than the slim version of that face varied as a function of the baseline corruptibility rating of the original photo, as planned in our preregistration. The ratings for each official in each version of the photo were first averaged over participants, and then these aggregate-level ratings for different photos of the same official were used to calculate perceptual biases. We did not observe significant correlation between perceptual bias (fat vs. slim versions of the photo) and the corruptibility inferences based on the original version of the photo, $\rho = 0.01$, $SE = 0.03$, 95% CI = [–0.06, 0.07], $p = 0.789$ (photo characteristics were included as control variables; see Fig. S8 in the Supplemental Material).

3.6 Discussion

Across three preregistered studies, we found evidence supporting the hypothesis that trait-specific inferences, such as corruptibility, made from photographs of officials' faces are associated with real-world measures of political corruption and violation. This association was replicated across officials at different levels of government. It was not driven by just a

small subset of faces or fully explained by other photo characteristics, such as smile intensity. The association remained robust when analyses controlled for heterogeneous beliefs about corruption base rates and potential photo-selection biases.

It is important to distinguish accuracy as defined by agreement with consensus judgments from accuracy related to actual real-world metrics (Funder, 1987). Similar to prominent studies of the association between competence judgments and election success (e.g., Todorov et al., 2005), our present work has pursued the latter interpretation of accuracy. The accuracy related to corruption records we found was comparable with that related to election success—for instance, Todorov et al. (2005) found that for 2004 U.S. Senate races, aggregate-level accuracy was 68.8%, and average individual-level accuracy was 53%. We emphasize that for our present work and a large literature on the association between face-based trait judgments and real-world metrics, accuracies at an individual level were only slightly above chance (but significantly so), and participants were very often wrong. However, the considerably larger effect sizes for aggregated judgments have important implications for real-world collective decisions such as elections and corruption investigations.

In Study 4, we found that an official was perceived as more corruptible when his face was manipulated to be slightly wider and less corruptible when his face was manipulated to be slightly slimmer, even though participants did not detect such manipulation of the facial identity. Our finding dovetails with the large literature on perceptual biases related to face width-to- height ratio (e.g., Deska et al., 2018) and the literature on weight stereotypes, which shows that overweight individuals are judged as lazy, greedy, selfish, and less trustworthy (Greenleaf, Chambliss, Rhea, Martin, & Morrow, 2006; Larkin & Pines, 1979).

Yet widening or narrowing the face potentially introduces other changes to the geometry of the face. It will be important for future studies to investigate which of these correlated structural changes are in fact detected by the brain and drive the change in social judgments that perceivers make.

The detailed causal mechanisms that ultimately underlie the association between a record of corruption and face-based judgments of corruptibility we found are likely to be complex and bidirectional (Swann, 1984). In particular, people who look corruptible might be more likely to be approached by others with the intent to corrupt them, which in turn results in the mutual behaviors required for corruption to occur (Kruglanski, 1989); further experimental studies would be required to tease apart their relative contributions.

Given these considerations, we emphasize that our findings should be interpreted with caution. Do they show that corruptible individuals have a different facial structure, as suggested by physiognomy? There are strong reasons to be skeptical. First, the record of an official is unlikely to be an errorless measure of how corruptible he actually is. Second, the photographs posted on government and campaign websites might provide a biased representation of an official's face—for example, some photos have clearly been retouched for skin texture and lighting. Third, there might well be other unknown confounding effects. For example, perhaps officials who committed a corrupt act might have looked more guilty when they posed for photos (even prior to being convicted), which in turn could have provided subtle visual cues for trait judgments. Fourth, both face judgments and social behaviors strongly depend on context (Todorov, 2017). For instance, in the context of business corruption, business executives' corruption records were not associated with how trustworthy they looked (Rule, Krendl, Ivcevic, & Ambady, 2013). These findings and the

considerations mentioned previously suggest that the self-fulfilling prophecy (e.g., Haselhuhn et al., 2013; Slepian & Ames, 2016) together with biases in judicial decisions (e.g., Wilson & Rule, 2015; Zebrowitz & McDonald, 1991) may be more plausible explanations than physiognomy for the face–corruption-record association we found. Future studies should examine multiple photographs of the same official taken in different contexts (e.g., posed and candid photos) and at different time points (e.g., from the time he ran for office, after misconducts, and after prosecutions).

There are important limitations to the generalizability of our studies (Simons, Shoda, & Lindsay, 2017). They leave open whether such trait judgments operate in the real world, where faces are not photographs and whether similar associations would hold for other cultures or for antisocial behaviors among people in general. It is also possible that corruptibility judgments are better correlated with other social behaviors that we did not measure, which in turn provide an indirect link to recorded corruption—indeed, it is conceivable that prosecutors' decisions might be one such social behavior.

We conclude with a future direction suggested by this work. The ultimate explanation for the findings we report must reside in evolutionarily based or experience-based neural mechanisms in the brains of both the subjects making the social judgments and the officials engaging in the corrupt behaviors. For instance, there are already claims that individual differences in traits can be predicted from patterns of brain activity (Dubois, Galdi, Paul, & Adolphs, 2018; Finn et al., 2015), and there is a large literature showing that social inferences engage specific neural networks in the brains of perceivers (Spunt & Adolphs, 2017). Future studies using neuroimaging could help further uncover the causal mechanisms behind our findings.

3.7 Supplementary Information

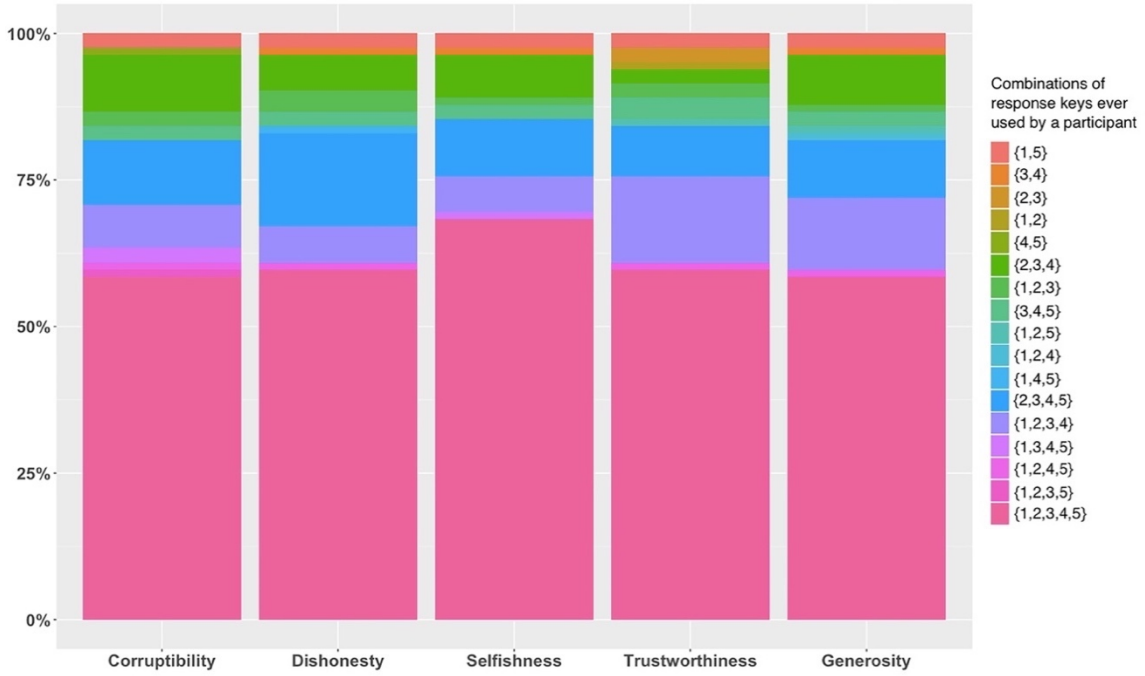


Fig. S1. Distributions of keys used by participants for trait judgments in Study 1 (N = 82). For the evaluation of each trait, the response keys, a participant had ever used to rate the faces were tracked. There are 31 possible combinations of response keys and 17 of them were observed in the current study.

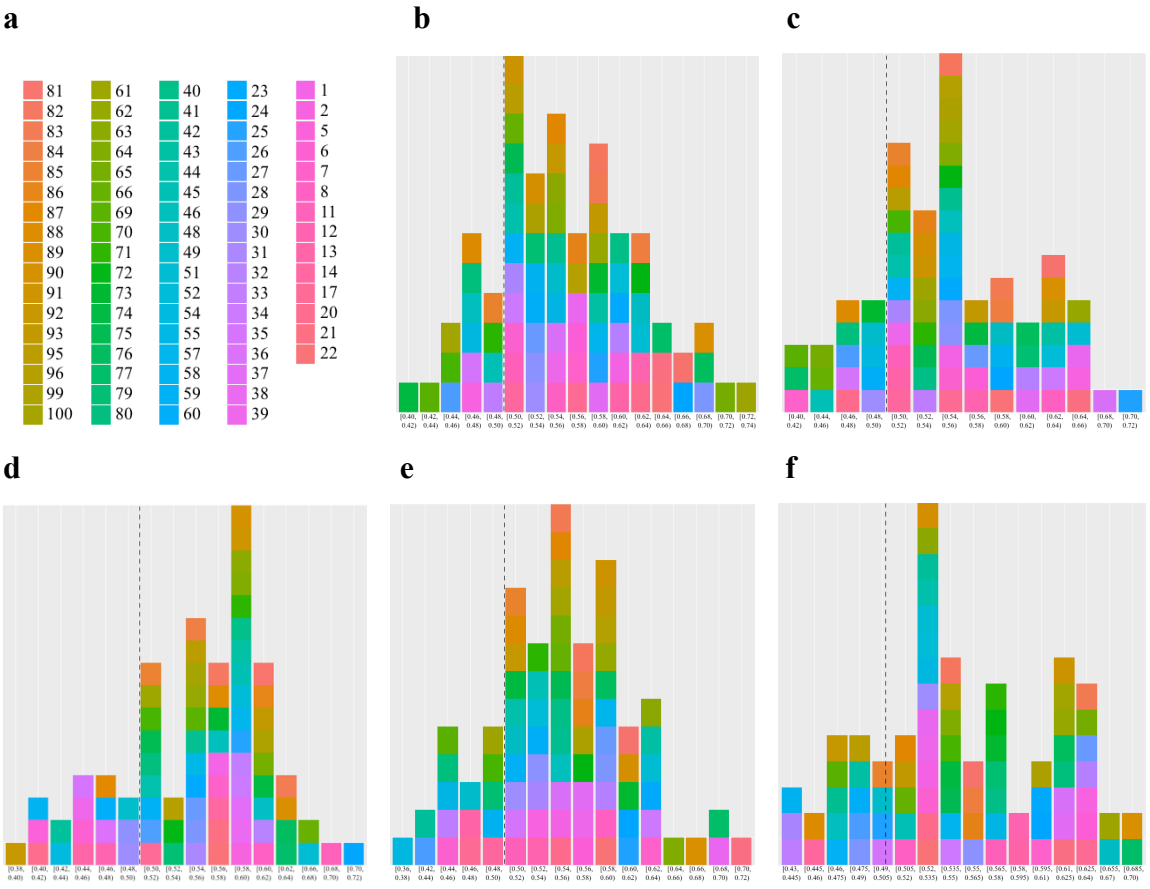
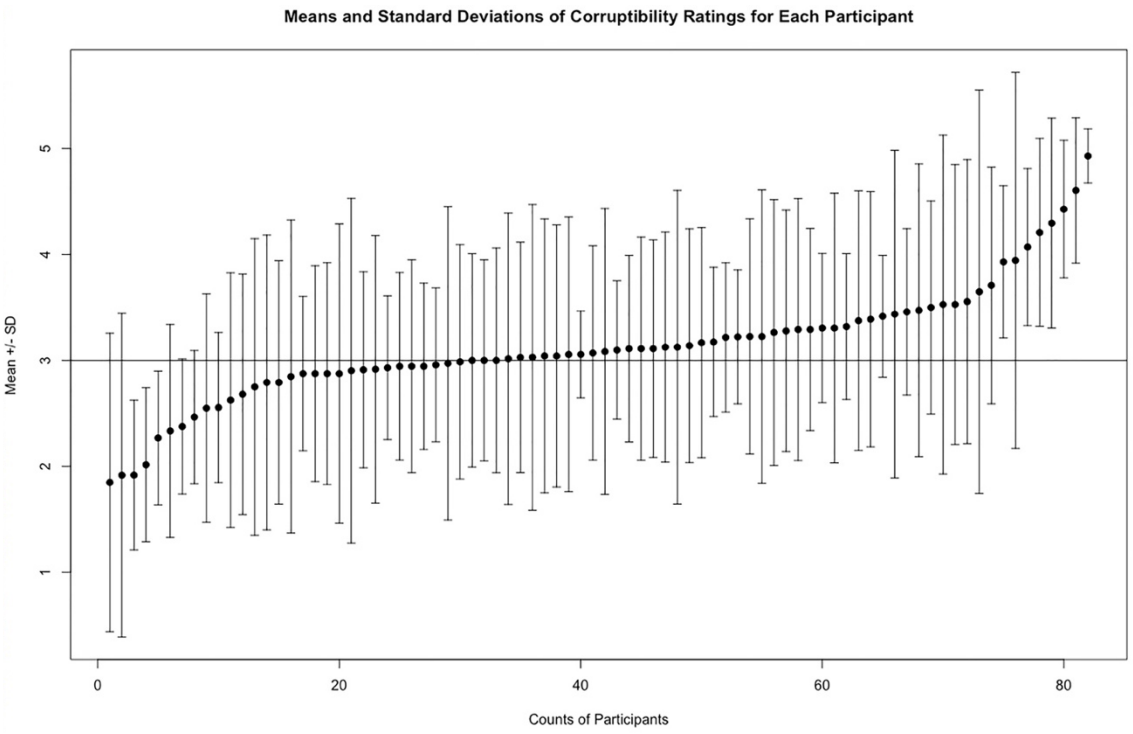
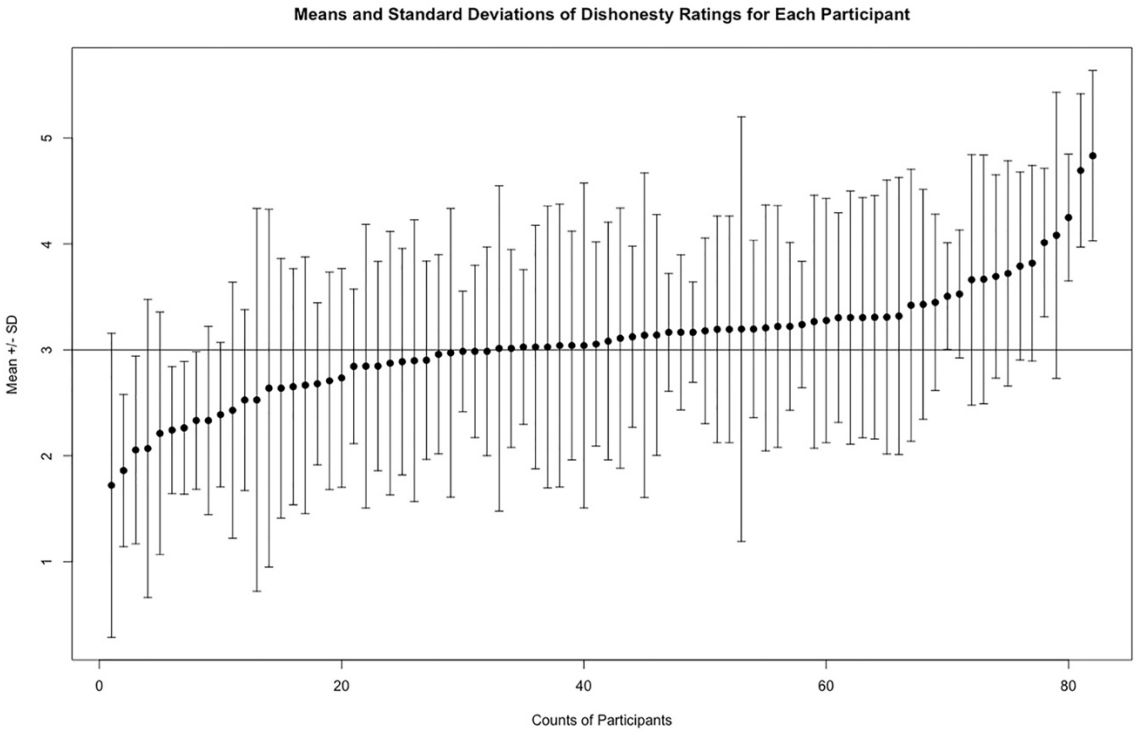


Fig. S2. Participant ID (N = 82) and corresponding color markers (a). The distribution of individual-level accuracies based on Corruptibility inferences (b). The dash-line indicates chance level accuracy (50%). The distribution of individual-level accuracies based on Dishonesty inferences (c). The distribution of individual-level accuracies based on Selfishness inferences (d). The distribution of individual-level accuracies based on Trustworthiness inferences (e). The distribution of individual-level accuracies based on Generosity inferences (f).

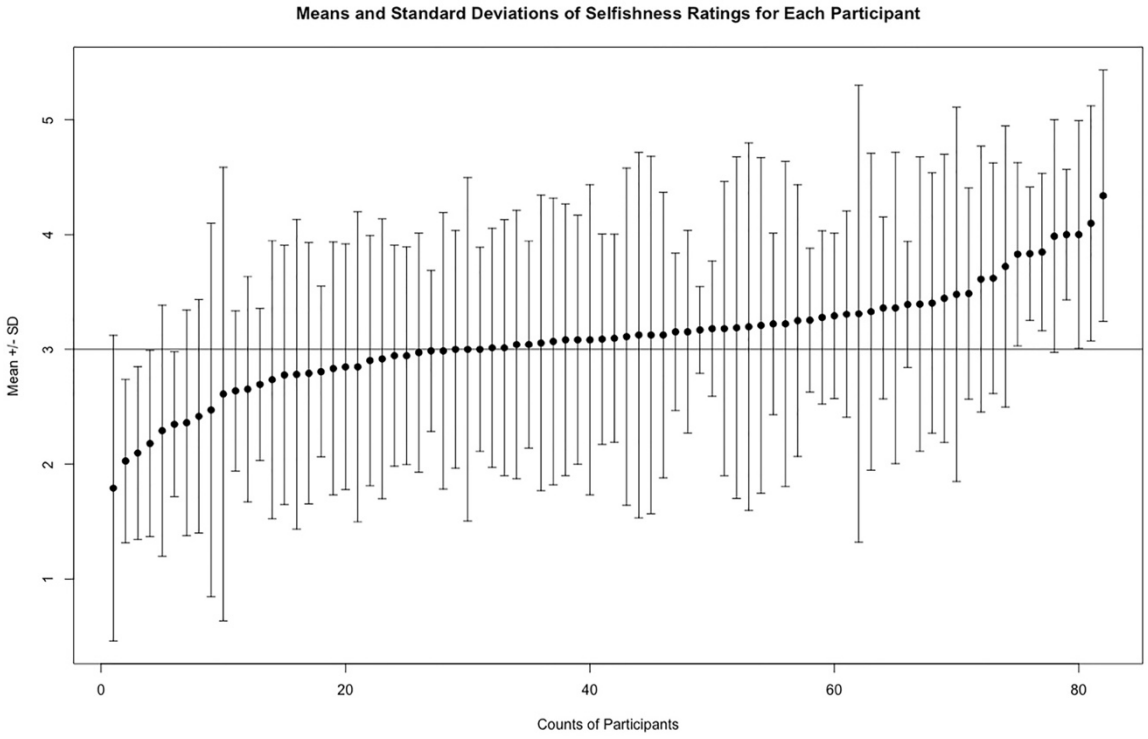
a


Means and Standard Deviations of Corruptibility Ratings for Each Participant
Mean +/- SD
1
2
3
4
5
0
20
40
60
80
Counts of Participants

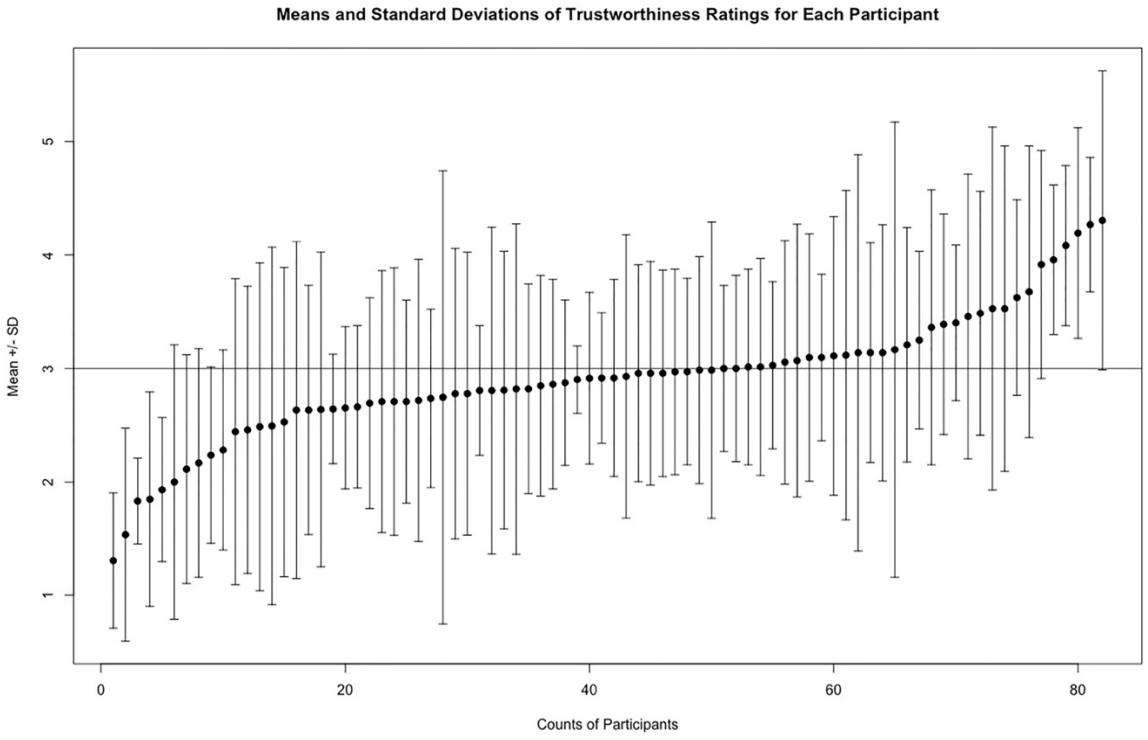
b


Means and Standard Deviations of Dishonesty Ratings for Each Participant
Mean +/- SD
1
2
3
4
5
0
20
40
60
80
Counts of Participants

c


Means and Standard Deviations of Selfishness Ratings for Each Participant
Mean +/- SD
1
2
3
4
5
0
20
40
60
80
Counts of Participants

d


Means and Standard Deviations of Trustworthiness Ratings for Each Participant
Mean +/- SD
1
2
3
4
5
0
20
40
60
80
Counts of Participants

e

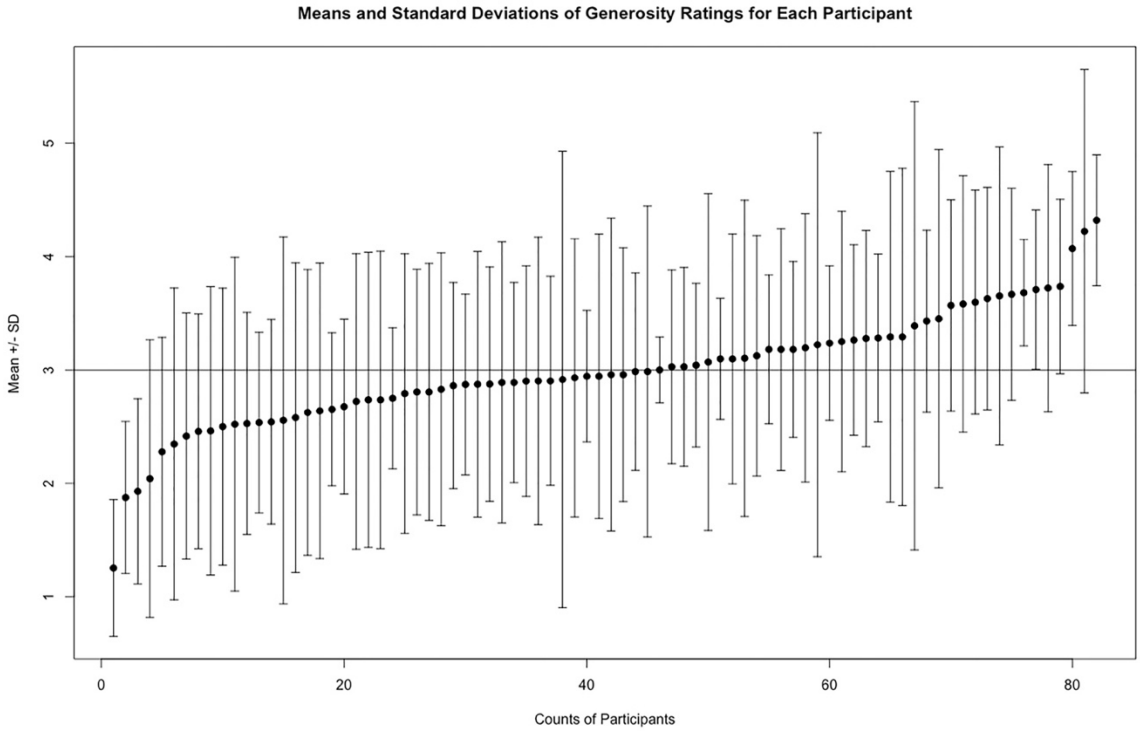


Fig. S3. Mean and standard deviations of trait judgments by each participant across faces for Corruptibility (a), Dishonesty (b), Selfishness (c), Trustworthiness (d), and Generosity (e) in Study 1.

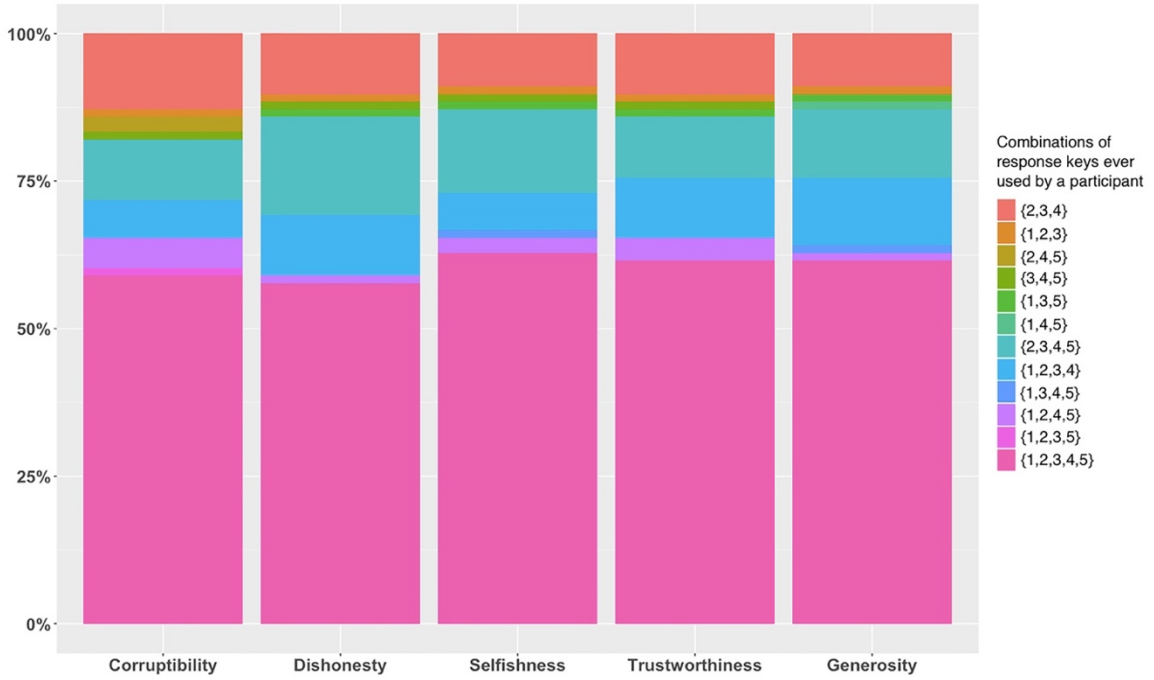


Fig. S4. Distributions of keys used by participants for trait judgments in Study 2 (N = 78). For the evaluation of each trait, the response keys, a participant had ever used to rate the faces were tracked. There are 31 possible combinations of response keys and 12 of them were observed in the current study.

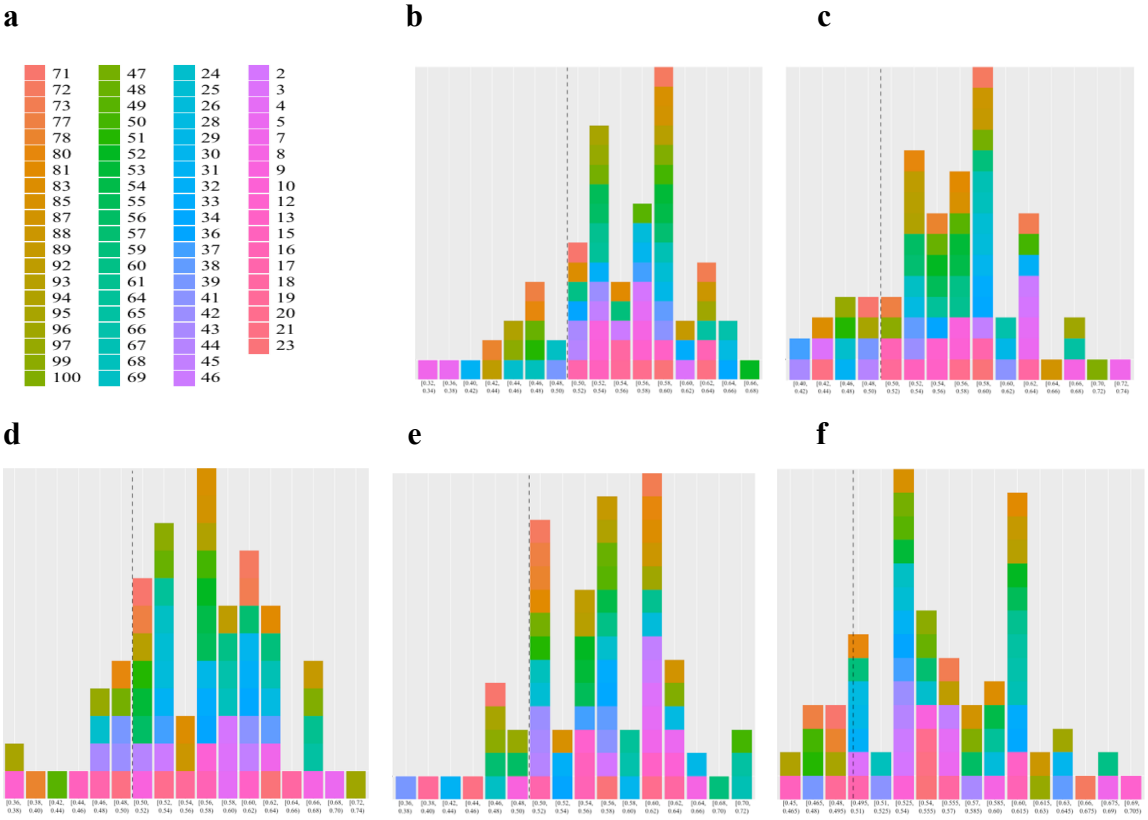
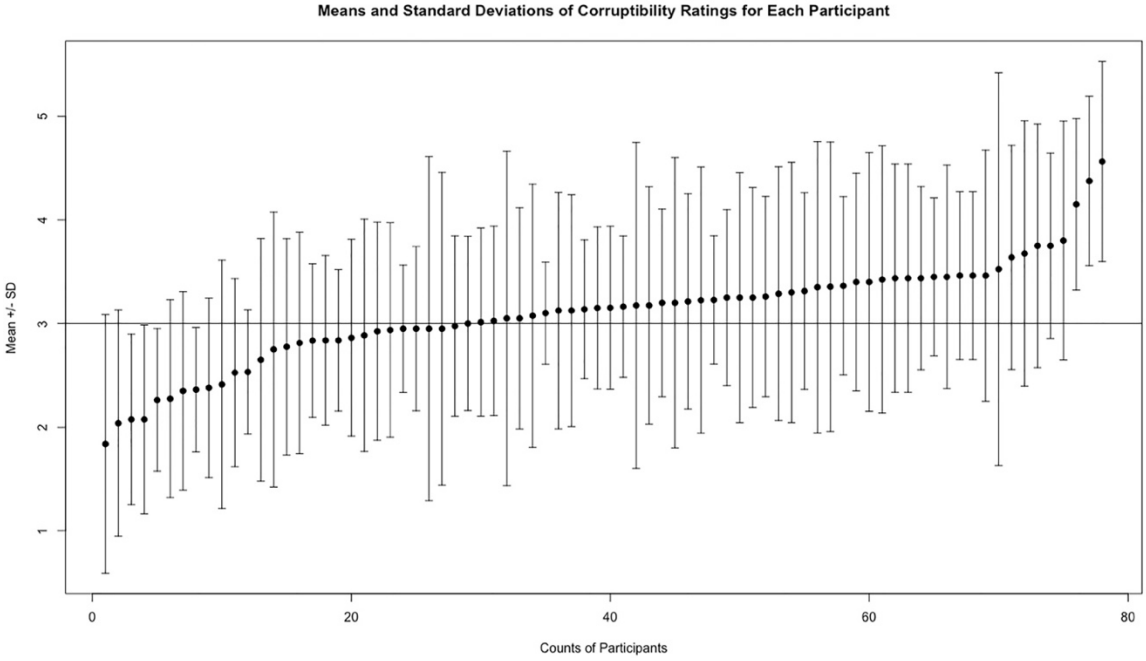
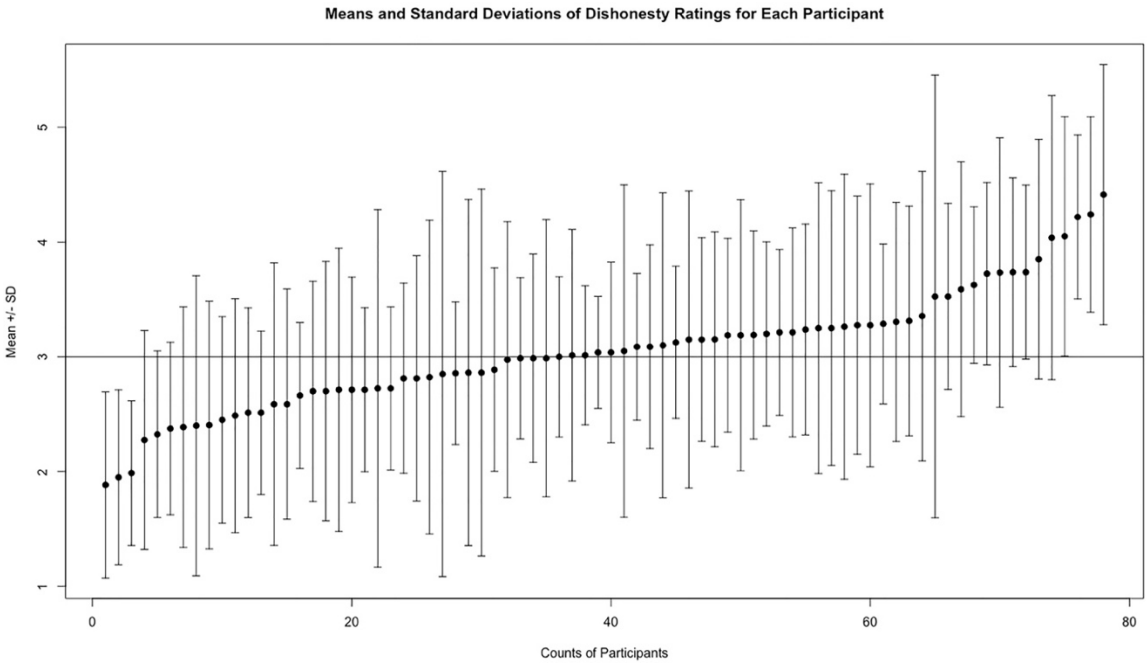


Fig. S5. Participant ID (N = 78) and corresponding color markers (a). The distribution of individual-level accuracies based on Corruptibility inferences (b). The dash-line indicates chance level accuracy (50%). The distribution of individual-level accuracies based on the Dishonesty inferences (c). The distribution of individual-level accuracies based on Selfishness inferences (d). The distribution of individual-level accuracies based on Trustworthiness inferences (e). The distribution of individual-level accuracies based on Generosity inferences (f).

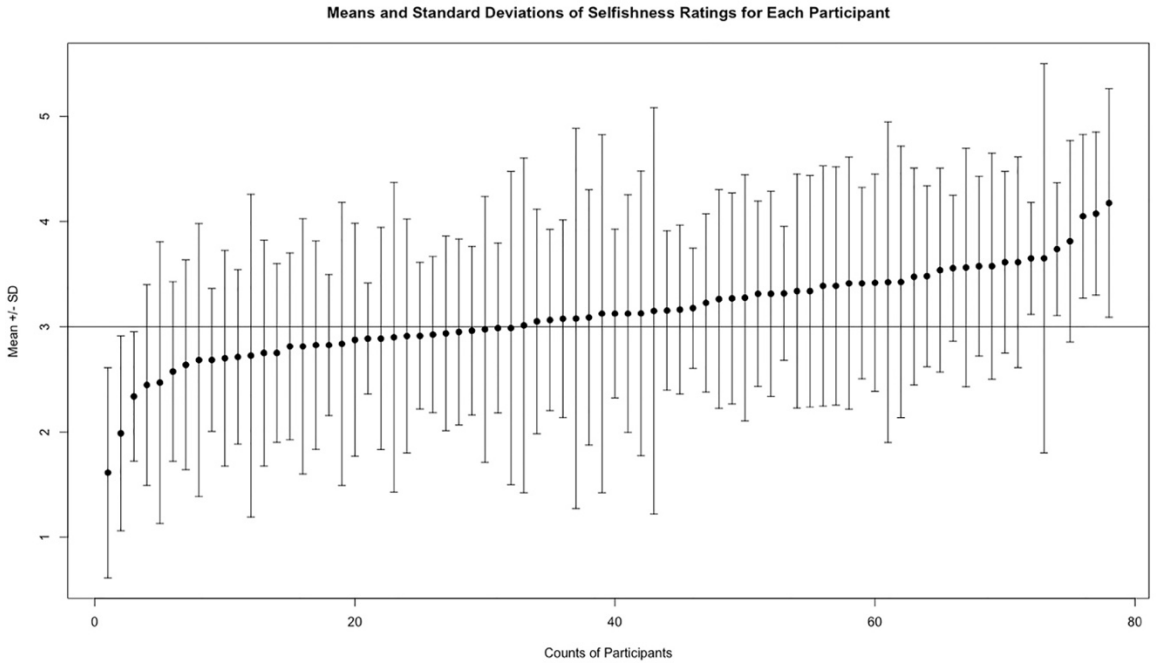
a


Means and Standard Deviations of Corruptibility Ratings for Each Participant
Mean +/- SD
1
2
3
4
5
0
20
40
60
80
Counts of Participants

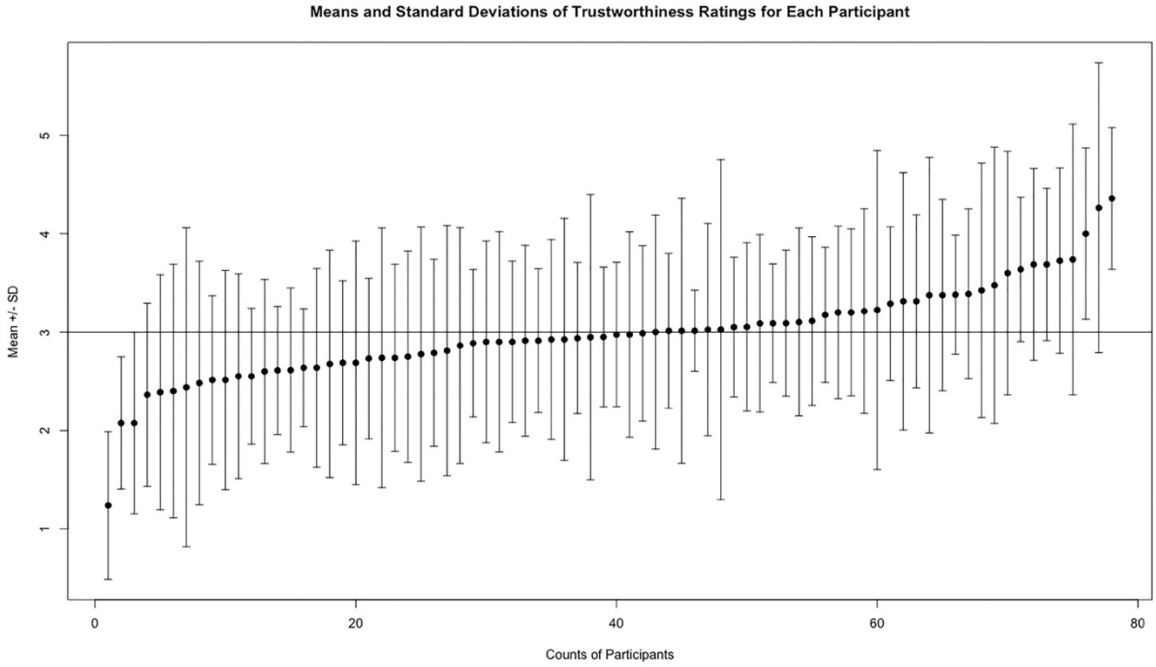
b


Means and Standard Deviations of Dishonesty Ratings for Each Participant
Mean +/- SD
1
2
3
4
5
0
20
40
60
80
Counts of Participants

c


Means and Standard Deviations of Selfishness Ratings for Each Participant
Mean +/- SD
1
2
3
4
5
0
20
40
60
80
Counts of Participants

d


Means and Standard Deviations of Trustworthiness Ratings for Each Participant
Mean +/- SD
1
2
3
4
5
0
20
40
60
80
Counts of Participants

e

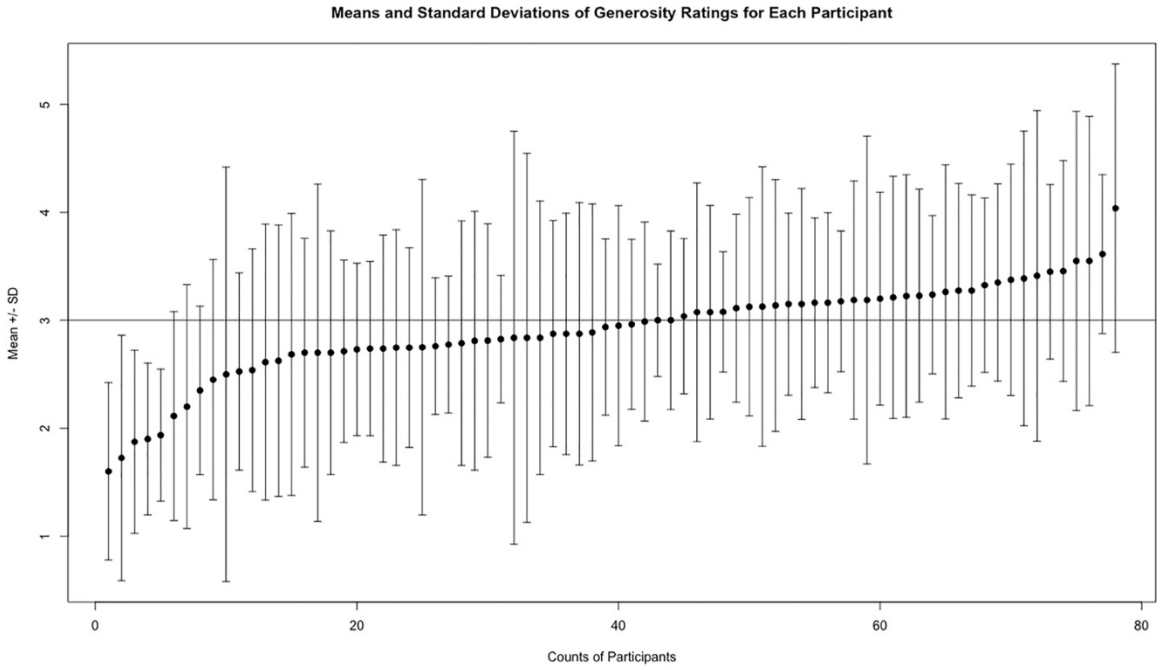


Fig. S6. Mean and standard deviations of trait judgments by each participant across faces for Corruptibility (a), Dishonesty (b), Selfishness (c), Trustworthiness (d), and Generosity (e) in Study 2.

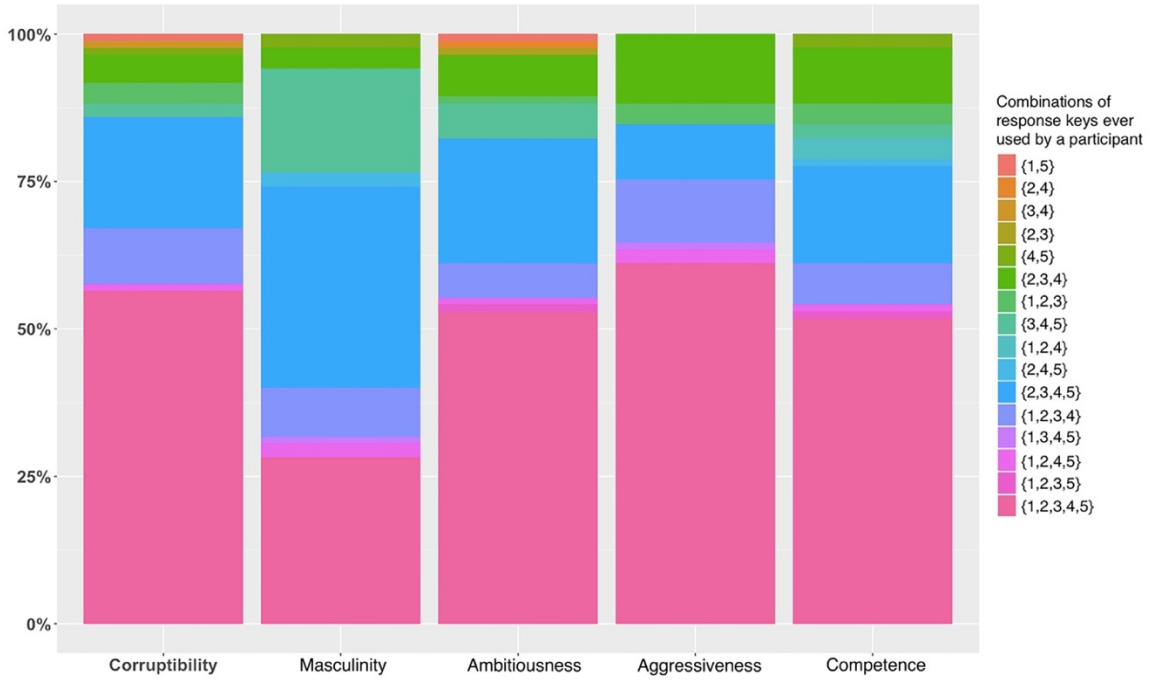


Fig. S7. Distributions of keys used by participants for trait judgments in Study 3 (N = 85). For the evaluation of each trait, the response keys a participant had ever used to rate the faces were tracked. There are 31 possible combinations of response keys, and 16 of them were observed in the current study.

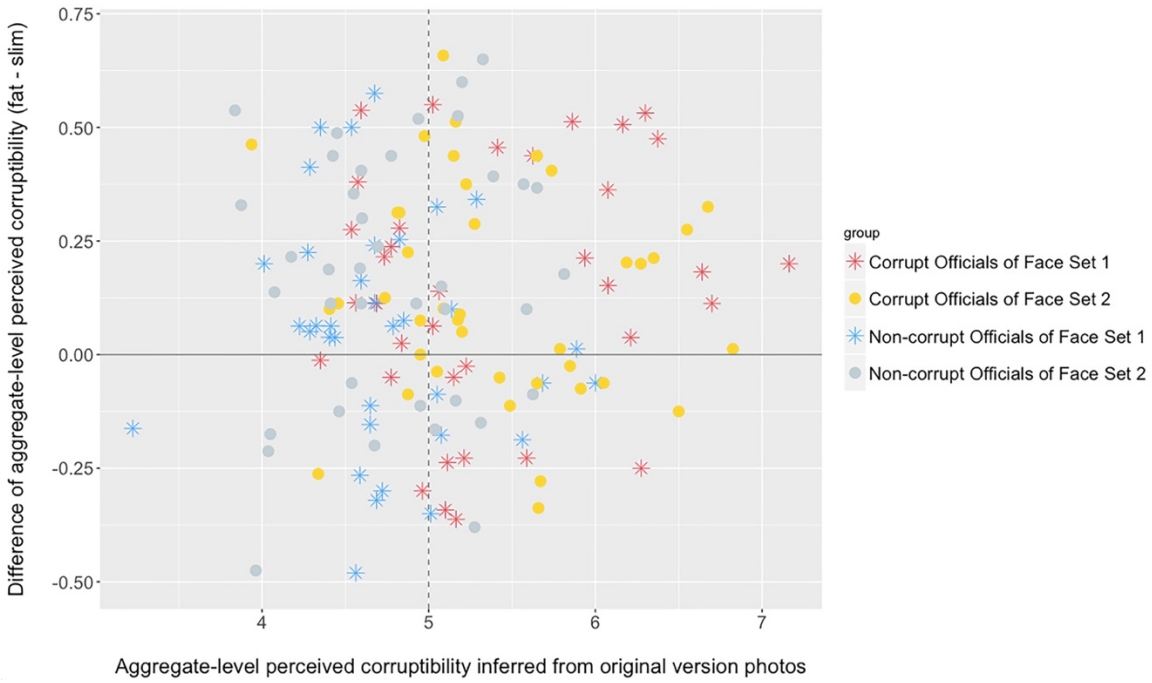


Fig. S8. Relation between aggregate-level corruptibility inferences based on the original-version photo and the perception difference between the fat- and slim-version photos for each elected official (N = 150). The vertical dashed line represents the midpoint of the rating scale, and the horizontal solid line indicates zero perception difference between the fat- and slim-version photos of the same official.

Table S1. Repeated measures correlations between each pair of traits calculated with individual-level ratings for Study 1 (N = 5757; N was determined by the number of participants multiplied by the number of faces excluding omitted observations; observations from a participant for a face would be omitted if ratings were not available for all the five traits).

	Corruptibility	Dishonesty	Selfishness	Trustworthiness
Dishonesty	0.25			
	[0.23, 0.28]			
Selfishness	0.31	0.24		
	[0.28, 0.33]	[0.22, 0.26]		
Trustworthiness	-0.29	-0.28	-0.30	
	[-0.31, -0.26]	[-0.30, -0.26]	[-0.33, -0.28]	
Generosity	-0.24	-0.21	-0.29	0.30
	[-0.27, -0.22]	[-0.24, -0.19]	[-0.31, -0.26]	[0.28, 0.32]

All p-values < 0.001.

Table S2. (Tie-corrected) Spearman correlation coefficients between each pair of traits calculated with aggregate-level ratings for Study 1 (N = 72).

	Corruptibility	Dishonesty	Selfishness	Trustworthiness
Dishonesty	0.88			
	[0.81, 0.92]			
Selfishness	0.84	0.85		
	[0.76, 0.90]	[0.77, 0.91]		
Trustworthiness	-0.84	-0.87	-0.83	
	[-0.90, -0.76]	[-0.92, -0.80]	[-0.89, -0.75]	
Generosity	-0.75	-0.83	-0.83	0.89
	[-0.84, -0.63]	[-0.89, -0.74]	[-0.89, -0.74]	[0.83, 0.93]

All p-values < 0.001.

Table S3. Percentages of Correctly Categorized Officials Based on Individual-level Trait Inferences from Study 1 with categorizing midpoint 3 in an alternative way.

	Average Individual-level Accuracy				
	Corruptibility	Dishonesty	Selfishness	Trustworthiness	Generosity
Mean Accuracy ($N = 82$)	53.51%	54.82%	54.21%	54.28%	54.70%
SD	6.24%	6.41%	6.92%	5.38%	5.93%
Lower Bound of 95% CI	52.37%	52.97%	52.94%	53.29%	53.61%
t-value ($df = 81$)	5.10	6.16	5.51	7.20	7.19
Cohen's d	0.56	0.68	0.61	0.80	0.79

For negative traits, a trial was accurate if the official was convicted of corruption and received a high (3, 4, or 5) rating from a participant, or, conversely, if he had a clean record and received a low (1 or 2) rating from a participant; for positive traits, a trial was accurate if the official was convicted of corruption and received a low (1, 2, or 3) rating from a participant, or, conversely, if he had a clean record and received a high (4 or 5) rating from a participant. All p-values < .001.

Table S4. Average individual-level accuracy calculated for subsets of stimuli in which the officials were excluded one by one following the order of the ranking (and one-sided t-tests against chance level) for Study 1.

	Corruptibility	Dishonesty	Selfishness	Trustworthiness	Generosity
Before Exclusion ($N = 72$)	55.73% ***	54.82% ***	55.10% ***	55.03% ***	54.97% ***
Excluded 1st	55.34% ***	54.44% ***	54.67% ***	54.64% ***	54.55% ***
Excluded 1st-2nd	55.01% ***	54.12% ***	54.31% ***	54.23% ***	54.16% ***
Excluded 1st-3rd	54.71% ***	53.81% ***	53.93% ***	53.87% ***	53.80% ***
Excluded 1st-4th	54.39% ***	53.49% ***	53.56% ***	53.50% ***	53.35% ***
Excluded 1st-5th	54.05% ***	53.18% ***	53.18% ***	53.14% ***	52.94% ***
Excluded 1st-6th	53.71% ***	52.82% ***	52.86% ***	52.82% ***	52.59% ***
Excluded 1st-7th	53.35% ***	52.51% **	52.56% ***	52.45% ***	52.20% *
Excluded 1st-8th	53.00% ***	52.21% **	52.24% **	52.08% **	51.77% *
Excluded 1st-9th	52.66% **	51.86% **	51.94% **	51.76% *	51.39% *
Excluded 1st-10th	52.35% **	51.47% *	51.62% *	51.45% *	51.00%
Excluded 1st-11th	52.04% *	51.09%	51.25%	51.13%	50.68%
Excluded 1st-12th	51.74% *	50.72%	50.91%	50.84%	50.33%
Excluded 1st-13th	51.44%	50.37%	50.60%	50.52%	49.97%
Excluded 1st-14th	51.12%	50.04%	50.28%	50.22%	49.63%

One-sample one-sided t-tests against chance (50%) were performed on the individual-level accuracies across participants for each exclusion. Signif. codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table S5. Coefficients and standard errors of general linear mixed model analyses on the association between officials' corruption records and inferences of each trait for Study 1 (*N* = 5757; *N* was determined by the number of participants times the number of faces minus omitted observations; observations from a participant for a face would be omitted if ratings were not available for all the five traits).

	Corruptibility	Dishonesty	Selfishness	Trustworthiness	Generosity
Trait Rating[a]	0.23 ***	0.17 ***	0.20 ***	-0.19 ***	-0.20 ***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Age	-0.02	-0.01	-0.01	-0.02	-0.02
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Glasses[b]	0.03	0.02	0.02	0.02	0.02
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
Bald[c]	-0.48 *	-0.47 *	-0.44 *	-0.46 *	-0.44 *
	(0.20)	(0.20)	(0.20)	(0.20)	(0.20)
Beard[d]	-0.05	-0.06	-0.07	-0.07	-0.06
	(0.19)	(0.19)	(0.19)	(0.19)	(0.19)
Mustache[e]	2.08 ***	2.05 ***	2.08 ***	2.06 ***	2.06 ***
	(0.13)	(0.13)	(0.13)	(0.13)	(0.13)
Smile Intensity[f]	-0.64 ***	-0.63 ***	-0.63 ***	-0.63 ***	-0.63 ***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Image Megapixels	-0.18 ***	-0.18 ***	-0.18 ***	-0.18 ***	-0.19 ***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Image Source: Wiki[g]	1.56 ***	1.56 ***	1.57 ***	1.57 ***	1.57 ***
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)
Image Source: News	-0.61 ***	-0.62 ***	-0.62 ***	-0.61 ***	-0.61 ***
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)

[a]Officials' corruption records were regressed on ratings of each trait respectively, presented in each column. [b]Glasses is a dummy variable with 1 indicating the official wore glasses. [c]Bald Head is a dummy variable with 1 indicating the official was bald headed. [d]Beard is a dummy variable with 1 indicating the official had a beard. [e]Mustache is a dummy variable with 1 indicating the official had a mustache. [f]Smile Intensity was coded manually with three levels (1 = smile with no teeth exposed, 2 = smile with teeth but not gums exposed, 3 = smile with gums exposed). [g]There were three sources of photos: government/campaign

websites (benchmark), Wikipedia, and news articles. All continuous variables were standardized. Signif. codes: *** p < 0.001, ** p < 0.01, * p < 0.05.

Table S6. Repeated measures correlations between each pair of traits calculated with individual-level ratings for Study 2 (N = 6115; N was determined by the number of participants multiplied by the number of faces excluding omitted observations; observations from a participant for a face would be omitted if ratings were not available for all the five traits).

	Corruptibility	Dishonesty	Selfishness	Trustworthiness
Dishonesty	0.31			
	[0.29, 0.34]			
Selfishness	0.26	0.35		
	[0.24, 0.29]	[0.32, 0.37]		
Trustworthiness	-0.31	-0.38	-0.33	
	[-0.33, -0.28]	[-0.40, -0.36]	[-0.35, -0.31]	
Generosity	-0.26	-0.33	-0.32	0.32
	[-0.28, -0.24]	[-0.35, -0.31]	[-0.34, -0.30]	[0.30, 0.34]

All p-values < 0.001.

Table S7. (Tie-corrected) Spearman correlation coefficients between each pair of traits calculated with aggregate-level ratings for Study 2 (N = 80).

	Corruptibility	Dishonesty	Selfishness	Trustworthiness
Dishonesty	0.88			
	[0.82, 0.92]			
Selfishness	0.85	0.91		
	[0.77, 0.90]	[0.86, 0.94]		
Trustworthiness	-0.89	-0.90	-0.91	
	[-0.93, -0.83]	[-0.94, -0.85]	[-0.94, -0.86]	
Generosity	-0.77	-0.84	-0.89	0.88
	[-0.85, -0.66]	[-0.90, -0.77]	[-0.93, -0.83]	[0.83, 0.92]

All p-values < 0.001.

Table S8. Percentages of Correctly Categorized Officials Based on Individual-level Trait Inferences from Study 2 with categorizing midpoint 3 in an alternative way.

	Average Individual-level Accuracy				
	Corruptibility	Dishonesty	Selfishness	Trustworthiness	Generosity
Mean Accuracy ($N = 78$)	53.94%	55.01%	54.56%	54.40%	54.77%
SD	6.34%	6.54%	6.16%	6.88%	6.09%
Lower Bound of 95% CI	52.74%	53.77%	53.40%	53.10%	53.63%
t-value ($df = 77$)	5.49	6.76	6.54	5.65	6.92
Cohen's d	0.62	0.77	0.74	0.64	0.78

For negative traits, a trial was accurate if the official was convicted of corruption and received a high (3, 4, or 5) rating from a participant, or, conversely, if he had a clean record and received a low (1 or 2) rating from a participant; for positive traits, a trial was accurate if the official was convicted of corruption and received a low (1, 2, or 3) rating from a participant, or, conversely, if he had a clean record and received a high (4 or 5) rating from a participant. All p-values < .001.

Table S9. Average individual-level accuracy calculated for subsets of stimuli in which the officials were excluded one by one following the order of the ranking (and one-sided t-tests against chance level) for Study 2.

	Corruptibility	Dishonesty	Selfishness	Trustworthiness	Generosity
Before Exclusion (N = 80)	54.72% ***	56.15% ***	55.78% ***	56.00% ***	55.80% ***
Excluded 1st	54.36% ***	55.73% ***	55.42% ***	55.53% ***	55.45% ***
Excluded 1st-2nd	54.02% ***	55.32% ***	55.10% ***	55.15% ***	55.09% ***
Excluded 1st-3rd	53.68% ***	54.94% ***	54.80% ***	54.78% ***	54.74% ***
Excluded 1st-4th	53.36% ***	54.55% ***	54.50% ***	54.40% ***	54.38% ***
Excluded 1st-5th	53.05% ***	54.20% ***	54.20% ***	54.01% ***	54.05% ***
Excluded 1st-6th	52.75% ***	53.84% ***	53.90% ***	53.63% ***	53.70% ***
Excluded 1st-7th	52.44% **	53.51% ***	53.60% ***	53.25% ***	53.35% ***
Excluded 1st-8th	52.14% **	53.17% ***	53.31% ***	52.87% ***	53.01% ***
Excluded 1st-9th	51.82% *	52.82% ***	53.01% ***	52.47% **	52.65% ***
Excluded 1st-10th	51.51% *	52.48% **	52.71% ***	52.10% **	52.30% ***
Excluded 1st-11th	51.20%	52.12% **	52.42% **	51.72% *	51.94% **
Excluded 1st-12th	50.87%	51.76% *	52.12% **	51.32%	51.57% **
Excluded 1st-13th	50.56%	51.38%	51.82% *	50.94%	51.26% *
Excluded 1st-14th	50.23%	51.00%	51.53% *	50.57%	50.97%
Excluded 1st-15th	49.90%	50.62%	51.25%	50.18%	50.69%
Excluded 1st-16th	49.56%	50.23%	50.97%	49.81%	50.39%
Excluded 1st-17th	49.22%	49.84%	50.67%	49.44%	50.09%

One-sample one-sided t-tests against chance (50%) were performed on the individual-level accuracies across participants for each exclusion. Signif. codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table S10. Coefficients and standard errors of general linear mixed model analyses on the association between officials' violation records and inferences of each trait for Study 2 (*N* = 6115; *N* was determined by the number of participants times the number of faces minus omitted observations; observations from a participant for a face would be omitted if ratings were not available for all the five traits).

	Corruptibility	Dishonesty	Selfishness	Trustworthiness	Generosity
Trait Rating[a]	0.24 ***	0.28 ***	0.27 ***	-0.26 ***	-0.27 ***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Age	0.12 ***	0.12 ***	0.12 ***	0.12 ***	0.12 ***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Glasses[b]	-2.61 ***	-2.59 ***	-2.62 ***	-2.61 ***	-2.62 ***
	(0.12)	(0.12)	(0.12)	(0.12)	(0.12)
Bald[c]	1.55 ***	1.55 ***	1.55 ***	1.55 ***	1.53 ***
	(0.15)	(0.15)	(0.15)	(0.15)	(0.15)
Beard[d]	-0.14	-0.15	-0.15	-0.14	-0.15
	(0.15)	(0.15)	(0.15)	(0.15)	(0.15)
Mustache[e]	1.48 ***	1.51 ***	1.49 ***	1.47 ***	1.49 ***
	(0.13)	(0.13)	(0.13)	(0.13)	(0.13)
Smile Intensity[f]	-0.30 ***	-0.28 ***	-0.29 ***	-0.26 ***	-0.26 ***
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
Image Megapixels	0.03	0.03	0.04	0.03	0.04
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Image Source: Gov[g]	-0.63 ***	-0.63 ***	-0.64 ***	-0.63 ***	-0.63 ***
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)

[a]Officials' violation records were regressed on ratings of each trait respectively, presented in each column. [b]Glasses is a dummy variable with 1 indicating the official wore glasses. [c]Bald Head is a dummy variable with 1 indicating the official was bald headed. [d]Beard is a dummy variable with 1 indicating the official had a beard. [e]Mustache is a dummy variable with 1 indicating the official had a mustache. [f]Smile Intensity was coded manually with two levels (0 = smile with no teeth exposed, 1 = smile with teeth exposed). [g]Image source was coded with two levels (1 = government/campaign websites, 0 = news articles). All continuous variables were standardized. Signif. codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table S11. Factor loadings of trait inferences on the first three factors identified in a principal components analysis with a Varimax rotation. The factor analysis was performed on the aggregate-level trait inferences.

	Factor Solution		
	Corruptibility-related	Competence-related	Masculinity-related
Corruptibility	**0.93**	-0.20	-0.08
Dishonesty	**0.93**	-0.22	-0.03
Selfishness	**0.93**	-0.16	-0.03
Trustworthiness	**-0.90**	0.35	-0.07
Generosity	**-0.87**	0.34	-0.12
Masculinity	0.09	0.20	**0.96**
Aggressiveness	**0.83**	0.10	0.44
Ambitiousness	-0.17	**0.96**	0.15
Competence	-0.52	**0.65**	0.39

Table S12. Summary statistics of facial structure metrics.

	Stimuli Set (n)	Mean	SD
Facial Width-to-Height Ratio	Set 1 (72)	2.21	0.22
	Set 2 (80)	2.26	0.23
Face Width/Lower Face Height	Set 1 (72)	1.29	0.11
	Set 2 (80)	1.29	0.12
Lower Face/Face Height	Set 1 (72)	0.58	0.05
	Set 2 (80)	0.58	0.03
Cheekbone Prominence	Set 1 (72)	1.06	0.05
	Set 2 (80)	1.04	0.04
Internal Eye Corner Distance	Set 1 (72)	0.24	0.05
	Set 2 (80)	0.24	0.03
Nose Height	Set 1 (72)	0.46	0.05
	Set 2 (80)	0.45	0.04
Mouth Width	Set 1 (72)	0.49	0.07
	Set 2 (80)	0.47	0.05
Nose/Mouth Width	Set 1 (72)	0.70	0.08
	Set 2 (80)	0.70	0.09

Table S13. Questions measuring whether participants noticed the width of the faces was manipulated.

Question	Format
1. Did you notice anything special about the photos in the experiment?	Open-ended
2. You might have noticed that photos of the same politician were shown for more than once in the experiment. Did you notice what are the differences among these photos of the same politician? Or do you think these photos of the same politician are identical?	Open-ended
3. In fact, the politicians' face-width has been manipulated and you have seen different versions of photos of the same politicians. Did you notice that the face of the same politician was wider in some photos and slimmer in others?	Closed-ended

3.8 References

1. Antonakis, J., & Eubanks, D. L. (2017). Looking leadership in the face. *Current Directions in Psychological Science*, *26*, 270–275.
2. Berry, D. S., & Zebrowitz-McArthur, L. (1988). What's in a face? Facial maturity and the attribution of legal responsibility. *Personality and Social Psychology Bulletin*, *14*, 23–33.
3. Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science*, *15*, 674–679.
4. 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.
5. Deska, J. C., Lloyd, E. P., & Hugenberg, K. (2018). Facing humanness: Facial width-to-height ratio predicts ascriptions of humanity. *Journal of Personality and Social Psychology*, *114*, 75–94.

6. Dubois, J. C., Galdi, P., Paul, L. K., & Adolphs, R. (2018). A distributed brain network predicts general intelligence from resting-state human neuroimaging data. *bioRxiv*, Article 257865. doi:10.1101/257865

7. Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience*, *19*, 1508–1519.

8. Farkas, L. G. (Ed.). (1994). *Anthropometry of the head and face*. New York, NY: Raven Press.

9. Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., . . . Constable, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, *18*, 1664–1671.

10. Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin*, *101*, 75–90.

11. Genevsky, A., & Knutson, B. (2015). Neural affective mechanisms predict market-level microlending. *Psychological Science*, *26*, 1411–1422.

12. Geniole, S. N., Denson, T. F., Dixson, B. J., Carré, J. M., & McCormick, C. M. (2015). Evidence from meta-analyses of the facial width-to-height ratio as an evolved cue of threat. *PLOS ONE*, *10*(7), Article e0132726. doi:10.1371/ journal.pone.0132726

13. Gheorghiu, A. I., Callan, M. J., & Skylark, W. J. (2017). Facial appearance affects science communication. *Proceedings of the National Academy of Sciences, USA*, *114*, 5970– 5975.

14. Greenleaf, C., Chambliss, H., Rhea, D. J., Martin, S. B., & Morrow, J. R. (2006). Weight stereotypes and behavioral intentions toward thin and fat peers among White and Hispanic adolescents. *Journal of Adolescent Health*, *39*, 546–552.

15. Hamermesh, D. S. (2011). *Beauty pays: Why attractive people are more successful*. Princeton, NJ: Princeton University Press.

16. Haselhuhn, M. P., Wong, E. M., & Ormiston, M. E. (2013). Self-fulfilling prophecies as a link between men's facial width-to-height ratio and behavior. *PLOS ONE*, *8*(8), Article e72259. doi:10.1371/journal.pone.0072259

17. Jussim, L. (1986). Self-fulfilling prophecies: A theoretical and integrative review. *Psychological Review*, *93*, 429–445.

18. Kruglanski, A. W. (1989). The psychology of being "right": The problem of accuracy in social perception and cognition. *Psychological Bulletin*, *106*, 395–409.

19. Larkin, J. C., & Pines, H. A. (1979). No fat persons need apply: Experimental studies of the overweight stereotype and hiring preference. *Sociology of Work and Occupations*, *6*, 312–327.

20. Lin, C., Adolphs, R., & Alvarez, R. M. (2017). Cultural effects on the association between election outcomes and face- based trait inferences. *PLOS ONE*, *12*(7), Article e0180837. doi:10.1371/journal.pone.0180837

21. Olivola, C. Y., Eastwick, P. W., Finkel, E. J., Hortacsu, A., Ariely, D., & Todorov, A. (2014). *A picture is worth a thou- sand inferences: First impressions and mate selection in Internet matchmaking and speed-dating*. Pittsburgh, PA: Tepper School of Business, Carnegie Mellon University.

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

23. Penton-Voak, I. S., Pound, N., Little, A. C., & Perrett, D. I. (2006). Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social Cognition*, *24*, 607–640.

24. Ravina, E. (2012). Love & loans: The effect of beauty and personal characteristics in credit markets. *SSRN*. doi:10.2139/ ssrn.1107307

25. 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*(3), Article e34293. doi:10.1371/ journal.pone.0034293

26. Rose-Ackerman, S. (2013). *Corruption: A study in political economy*. San Diego, CA: Academic Press.

27. Rule, N. O., Ambady, N., Adams, R. B., Jr., Ozono, H., Nakashima, S., Yoshikawa, S., & Watabe, M. (2010). Polling the face: Prediction and consensus across cultures. *Journal of Personality and Social Psychology*,

28. *98*, 1–15.

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013).

29. Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology*, *104*, 409–426.

30. Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123–1128.

31. Slepian, M. L., & Ames, D. R. (2016). Internalized impressions: The link between apparent facial trustworthiness and deceptive behavior is mediated by targets' expectations of how they will be judged. *Psychological Science*, *27*, 282–288.

32. Spunt, R. P., & Adolphs, R. (2017). The neuroscience of understanding the emotions of others. *Neuroscience Letters*. Advance online publication. doi:10.1016/j.neulet.2017.06.018

33. Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science*, *21*, 349–354.

34. Swann, W. B. (1984). Quest for accuracy in person perception: A matter of pragmatics. *Psychological Review*, *91*, 457–477.

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

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

37. 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.

38. Valentine, K. A., Li, N. P., Penke, L., & Perrett, D. I. (2014). Judging a man by the width of his face: The role of facial ratios and dominance in mate choice at speed-dating events. *Psychological Science*, *25*, 806–811.

39. Valla, J. M., Ceci, S. J., & Williams, W. M. (2011). The accuracy of inferences about criminality based on facial appearance. *Journal of Social, Evolutionary, and Cultural Psychology*, *5*(1), 66–91.

40. Van't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, *108*, 796–803.

41. Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science*, *26*, 1325–1331.

42. Zebrowitz, L. A., & McDonald, S. M. (1991). The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior*, *15*, 603–623.

Chapter 4

The comprehensive space for trait attributions from faces is four dimensional

Modern psychological models of person perception postulate that we attribute traits to people from their faces along two or three major dimensions. Yet, these models rely on studies that used only a small number of traits, possibly obscuring a higher dimensionality. Here, we applied deep neural networks to representatively sample multiple stimulus sets, and derived a novel set of 100 traits and 100 faces for a comprehensive protocol we administered in two pre-registered studies. Study 1 collected 750,000 sparse online ratings from 1,500 participants and found four dimensions: critical/condescending, leadership/competence, female-stereotype, and youth-stereotype. Study 2 collected complete datasets on-site from 210 participants (2,100,000 trials) in seven different countries, and largely reproduced this finding, even in single participants. Test-retest reliability and direct comparisons with other trait spaces from the literature provide the most comprehensive characterization of person perception from faces.

4.1 Introduction

Humans spontaneously attribute a wide range of traits to other people from merely seeing their faces, such as attributions of demographics (e.g., gender, age, sexual orientation), physical appearance (e.g., baby-faced, feminine, healthy), social evaluation (e.g., trustworthy, competent, dominant), and personality (e.g., aggressive, sociable, serious)[1–7]. Although we are often wrong about our attributions, they nonetheless have important consequences in real life: they influence decisions of who to trust and who to punish in laboratory experiments[8–10], and who to elect and who to jail in the real world[11–15], even though the diagnostic validity of trait attributions from faces remains inconclusive[16–22].

Despite a considerable amount of work on the topic, it remains unclear how to characterize the trait attributions that we make. Two disparate literatures have argued that trait attributions of people are well described by either five or three dimensions. The first literature concerns our understanding of personality—the relatively objective and stable dispositions of a person23. To tackle the challenge of adequately sampling attributes that span the entire domain of personality, lexical studies of personality posit that "the most important individual differences eventually become encoded as single words in the natural language"[23,24]. These studies extracted hundreds of personality trait-words from the English lexicon and used them to elicit self- and peer-evaluations of personality. These studies[25–27] generally support the popular Five-Factor Model (the Big Five) which theorizes that five dimensions capture individual differences in personality: agreeableness, conscientiousness, emotional stability, extraversion, and openness to experience[28–36].

The second large literature is based not on self- and peer-evaluations of trait-words but on judgements of unfamiliar faces—plausibly the most ubiquitous social stimuli to which we attribute such trait-words in everyday life. This literature has emphasized trait attributions concerning demographic, physical, and social characteristics of people, given that those traits are frequently and spontaneously inferred from faces[1,4,5,37,38]. Seminal work[1] using facial images as stimuli to prompt participants to attribute thirteen traits (including personality traits) to unfamiliar faces has shown that these evaluations fall along two dimensions: trustworthiness and dominance. A subsequent study analyzing facial inferences with a different set of thirteen traits revealed a third dimension, youthful/attractiveness[38]. Intriguingly, the three dimensions that have emerged from the face evaluation literature (trustworthiness, dominance, youthful/attractiveness) have all been categorized as social evaluative traits rather than personality traits according to the study of trait-words[23,27].

The discrepancy between the five-dimensional framework of personality theories, and the three-dimensional one from the face evaluation literature likely arises, at least in part, because both sets of findings are based on an incomplete set of traits. For instance, the lexical studies of personality excluded a large number of traits that are not commonly considered personality traits—attributions of demographic characteristics, physical appearance, social evaluation, and temporary states[23,24]. What is missing from both literatures is a list of traits across diverse categories that is as complete as possible, a criterion we believe is essential for uncovering the underlying dimensions of person perception from faces[39]. Needless to say, this is a major challenge: one would like to adequately sample traits that span the entire domain, but that are also meaningful, non-redundant, and can be reliably inferred from faces[40].

We set out to meet this challenge in the present project: to provide a comprehensive investigation on the dimensions that underlie person perception from faces by incorporating traits from all important categories (demographics, physical appearance, social evaluation, personality, emotion). To tackle the challenge of adequately sampling traits that span such a broad range, we drew upon previous personality and face evaluation studies to assemble an inclusive set of 482 traits, to which we applied natural language processing techniques together with other filters to derive a subset of 100 trait-words that sampled clear, non-redundant meanings in a representative manner (Fig. 1a and Methods). Aiming for a similarly well-sampled set of face stimuli, we also used a deep neural network to sample a set of 100 faces from a total of 426 gleaned across three face databases (Fig. 1b and Methods).

We pre-registered both of our studies and predetermined sample sizes based on the estimation of stable averaged measures recommended by a recent study[41]. Study 1 (see preregistration at https://osf.io/6p542/?view_only=fff024253b604edb832a9824cbdafa75) examined attributions of the one hundred traits in a large online sample (L = 750,000 sparse ratings and N = 1,500 participants; participants did not rate all faces on all traits). Study 2 (preregistration https://osf.io/qxgmw/?view_only=fd43b2e8b25248f7b7de51b9aeae1894) investigated the reproducibility and generalizability of the findings by collecting complete datasets (all faces rated on all traits by each participant) in seven countries and regions (North America, Latvia, Peru, the Philippines, India, Kenya, and Gaza; N = 210 participants). Several additional analyses document test-retest reliability and consensus of attributions across the traits; explore how the four dimensions found in the present research relate to the Big Five personality dimensions, the three face evaluation dimensions from the

literature, and the semantic dimensions of the trait-words; and inspect how the dimensions of person perception from faces uncovered from aggregate-level data (Study 1) relate to those uncovered from complete individual-level data (Study 2).

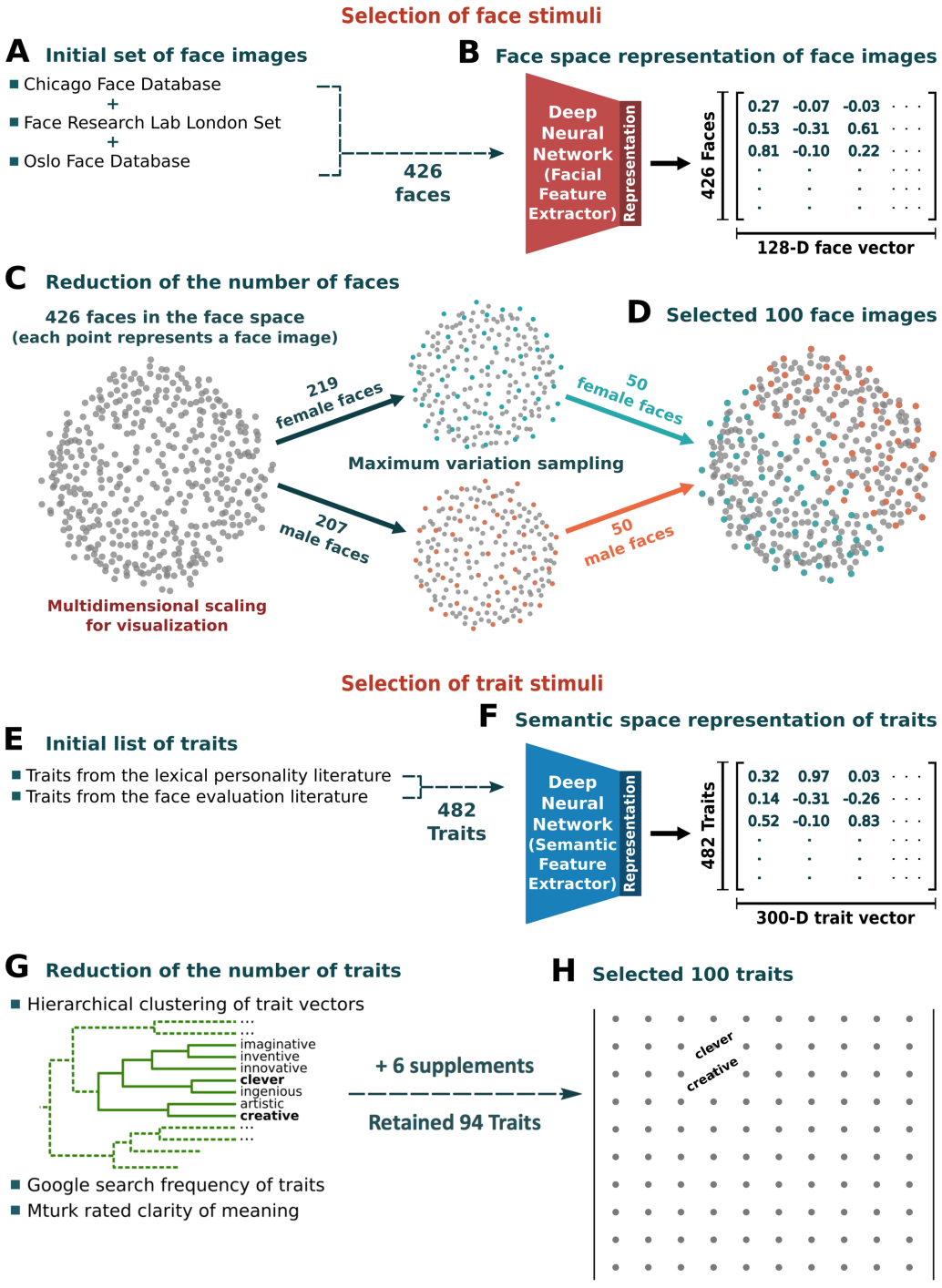


Fig. 1| Sampling methods for face stimuli (A-D) and trait words (E-H). (A) Our initial set of face stimuli included 426 Caucasian faces from three databases[42–44] that were frontal, with direct eye gaze and neutral facial expressions, and without glasses or other objects obscuring the face. (B) Each face was represented with a vector of 128 computer-extracted facial features (for recognizing facial identities) using a state-of-the-art neural network[45] that had been pretrained to identify individuals across millions of faces (of all different aspects and races). (C) Maximum variation sampling[46] was applied to select faces with maximum variability in facial structure, and a final set of 100 faces was obtained (D). (E) Our initial list of traits integrated personality traits, demographic traits, physical traits, social evaluative traits, and emotional traits from multiple areas of previous research[1,2,25,47]. (F) Each trait was represented with a vector of 300 computer-extracted semantic features (for word embeddings and text classification) using a state-of-the-art neural network[48] that had been pretrained to assign words to their contexts across 600 billion words. (G) We further filtered words to remove synonyms, antonyms, words with unclear meaning, or infrequent usage. (H) The sampled traits were supplemented with 4 demographic and health characteristics (education, income, sexual orientation, autism) and 2 derogatory words (idiot, loser) that are frequently used to describe a person in natural language (see Supplemental Methods for the complete list of traits and their definitions provided to participants).

4.2 Results

Study 1

Reliability and consensus of trait ratings

Participants viewed color images of the one hundred face stimuli (one at a time) and assigned ratings regarding a trait using a 7-point Likert scale (Methods). Following our preregistered data exclusion criteria (Methods), of the full sample with a registered size of $N = 1{,}500$ participants and $L = 750{,}000$ ratings (Methods), $n = 48$ participants and $l = 27{,}491$ ratings were excluded from further analysis. We first verified that the faces included in our

stimulus set produced sufficient variance in ratings as intended. The distribution of average ratings per face for each trait showed that the faces elicited a wide range of ratings for all traits (mean range across traits = 3 points on our 7-point Likert scale, excluding *white* and *feminine*; Supplementary Fig. 1). As expected, since all faces in the study were from white volunteers and were balanced in gender, almost all faces received high average ratings on *white* and bimodal ratings on *feminine*.

Next, we examined the within-subject test-retest reliability of trait ratings following our preregistered data analysis plan. Each of the one hundred traits was rated twice for all faces by nonoverlapping subsets of participants (ca. n = 15 per trait). We applied linear mixed-effect modeling (Methods) to handle the potential data non-independence due to repeated-measure designs (each participant provided multiple ratings), which adjusted for non-independence by incorporating both fixed effects (that were constant across participants) and random effects (that varied across participants). Ratings from every participant for every face collected at the second time were regressed on those collected at the first time (ca. l = 1,445 pairs of ratings per trait after data exclusion) while controlling for the random effect of participants. As hypothesized in our preregistration, we found that ratings of traits about physical appearance, such as *white* (r = 0.81), *feminine* (r = 0.80), *strong* (r = 0.68), *youthful* (r = 0.67), *baby-faced* (r = 0.67), *beautiful* (r = 0.67) had high within-subject test-retest reliabilities (Fig. 2). To our surprise, ratings of *autistic* also showed a high test-retest reliability (r = 0.64). Ratings of whether the person had low or high *income* showed the lowest test-retest reliability (r = 0.22). However, ratings of all the one hundred traits had acceptable test-retest reliabilities ($r > 0.20$)[1].

Before inspecting the between-subject consensus of trait ratings, we explored whether potential individual differences in the participants might account for different ratings of a trait. We focused on age, gender, education, and psychiatric or neurological illness (see Methods for details). For each of these four characteristics, we median-split participants into two groups and applied mixed-effect modeling to estimate the effect of each characteristic on the ratings of each trait (Methods). Ratings of every face by every participant (ca. $n = 58$ participants and $l = 5{,}780$ ratings per trait after data exclusion) were regressed on each characteristic while controlling for the random effects of individual participants. The distribution of effect sizes for the 400 regressions (100 traits by 4 characteristics) showed that none of the participant characteristics had a significant effect on trait ratings after Bonferroni correction (before Bonferroni correction, 18 of the 400 effects were significant; e.g., compared to younger participants, older participants gave the faces lower ratings on *strong, prudish, wise* and higher ratings on *healthy*; compared to participants with psychiatric or neurological illness, those without gave the faces lower ratings on *articulate, intellectual,* and *compulsive*; see Supplementary Fig. 2).

We analyzed the between-subject consensus of trait ratings following our preregistered data analysis plan, computing for each trait the intraclass correlation coefficient (ICC(2,k), see Methods), using ratings of every face by every participant (ca. $n = 58$ participants and $l = 5{,}780$ ratings per trait after data exclusion). A high intraclass correlation coefficient indicates that the total variance in the ratings is mainly explained by the rating variance across faces instead of variance across participants. We observed excellent between-subject consensus (ICCs greater than 0.75) for ninety-three of the one hundred traits (Fig. 2). Traits with the highest between-subject consensus were those concerning physical appearance,

such as *feminine*, *white*, *youthful*, *strong*, *beautiful*, and *baby-faced*. The seven traits with the lowest consensus (*self-critical*, *sarcastic*, *reserved*, *anxious*, *thrifty*, *shallow*, and *compulsive*) still had good ICCs (ICCs ranged from 0.60 to 0.75).

We further confirmed that between-subject consensus was attenuated by within-subject test-retest reliability (people cannot agree with one another any better than they can agree with themselves). To compare a trait's between-subject consensus with its within-subject reliability, we reassessed both quantities with the same metric, Spearman rank-order correlations. Results verified that the within-subject reliability of every trait was indeed always higher than its between-subject consensus (Supplementary Fig. 3).

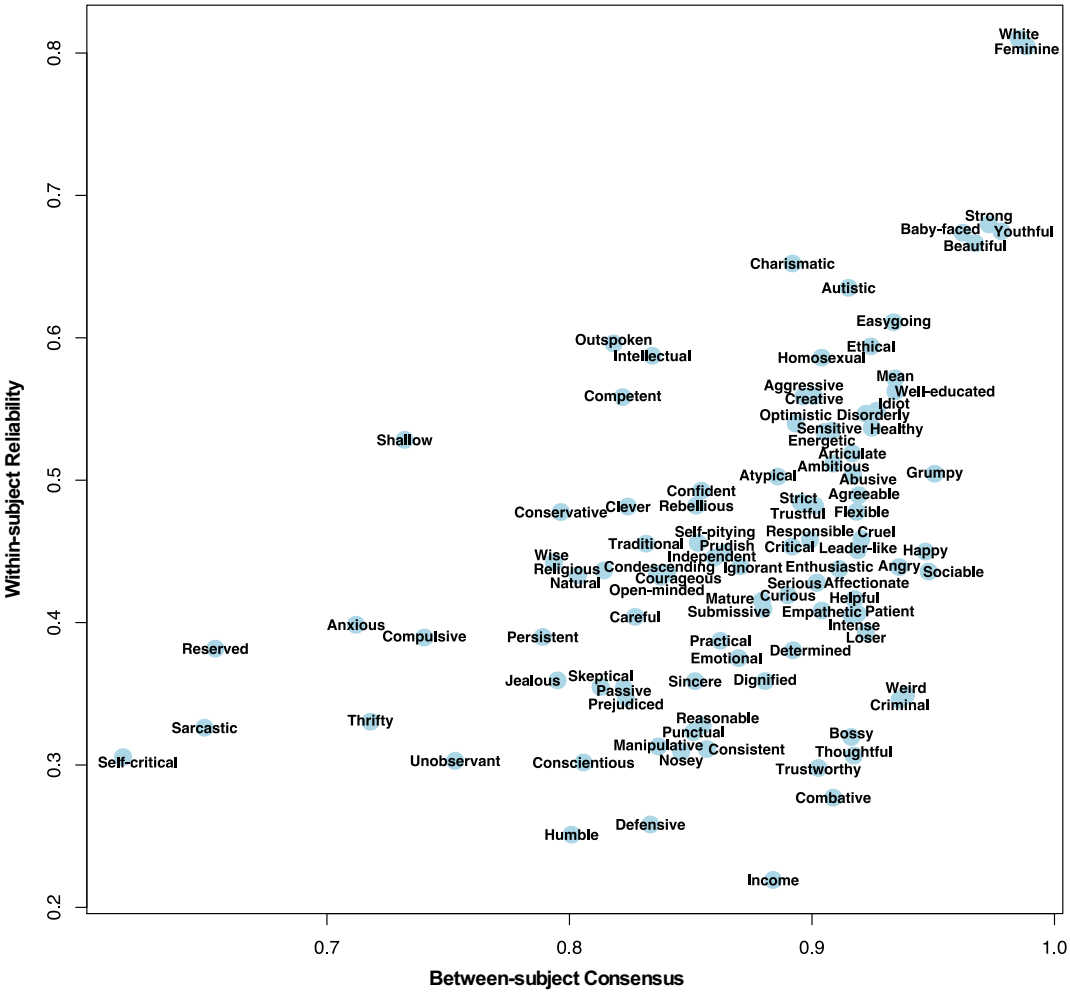


Fig. 2| Within-subject test-retest reliability and between-subject consensus for the attributions of one hundred traits from faces. The vertical axis indicates the within-subject test-retest reliability assessed with linear mixed-effect modeling. The horizontal axis indicates the between-subject consensus assessed with intraclass correlation coefficients. See Supplementary Fig. 3 for full details and the comparison between within-subject test-retest reliability and between-subject consensus for each trait.

Four dimensions characterize trait attributions from faces

We next analyzed how attributions of different traits were related to one another using aggregate ratings across participants. Since aggregate ratings were all on the same scale, we used Pearson correlations. Visual inspection of the correlation matrix suggested that most traits were highly or moderately correlated (either positively or negatively; Supplemental Figure 4a), but that eight traits (*thrifty*, *shallow*, *sarcastic*, *white*, *conservative*, *homosexual*, *nosey*, *reserved*) showed discontinuously low correlations with most other traits ($r < 0.3$; Supplementary Fig. 4b). Since this raises concerns for the factor analysis we wished to undertake next (i.e., low factorability), we excluded these eight traits from further analysis (including them did not substantially change the final results and still yielded 4 dimensions, see Supplementary Fig. 5a and 5b).

Following our preregistered data analysis plan, we applied exploratory factor analysis (EFA; see Methods) to examine the correlation structure among the remaining 92 strongly intercorrelated traits and to derive a small number of factors that represent their shared variance. We verified the sampling adequacy (Kaiser-Meyer-Olkin test of sampling adequacy, MSA = 0.79) and sphericity (Bartlett's $p < .001$) of the dataset (an *m by n* dimensional dataset with aggregate ratings for $m = 100$ faces by $n = 92$ traits). As

recommended[49–51], we applied parallel analysis to determine the optimal number of factors to retain in EFA. Parallel analysis retains factors that are not simply due to chance by comparing the eigenvalues of the observed data matrix with those of multiple randomly generated data matrices that match the sample size of the observed data matrix. This produces accurate estimations consistently across different conditions (e.g., the distribution properties of the data)[49–51]. For comparison, we also obtained estimations based on Kaiser's rule, Cattell's scree test, the acceleration factor, and the optimal coordinates index. Common factor parallel analysis with 5,000 Monte Carlo simulations showed that four factors explained the underlying structure of our dataset (Supplementary Fig. 5a). Cattell's scree test and the optimal coordinates index agreed with parallel analysis on a four-factor solution.

We thus performed EFA on the dataset to extract four factors, using the *MinRes* procedure to extract the factors, and the oblique rotation (*oblimin*) to transform the solutions (Methods). *MinRes* produces solutions that are very similar to maximum likelihood estimation and works even when the matrix is singular; oblique rotation is a more appropriate transformation than orthogonal rotation (e.g., *varimax*) in our case because it allows freedom for the factors to be correlated with or independent of each other. Results of the EFA showed that the four factors each explained 31%, 31%, 11%, and 12% of the common variance in the data, collectively accounting for 85% (if a five-factor solution was derived with the same approach, the five factors collectively accounted for 87% of the common variance in the data). An examination of the standardized loading of each trait on the four factors suggested an interpretation of the factors as: critical/condescending, leadership/competence, female-stereotype, and youth-stereotype (Fig. 3). Since oblique

rotation allowed factors to be correlated, they turned out to be weakly correlated in expected directions (r_{12} = -0.15, r_{13} = -0.33, r_{14} = -0.23, r_{23} = 0.21, r_{24} = 0.33, r_{34} = 0.12).

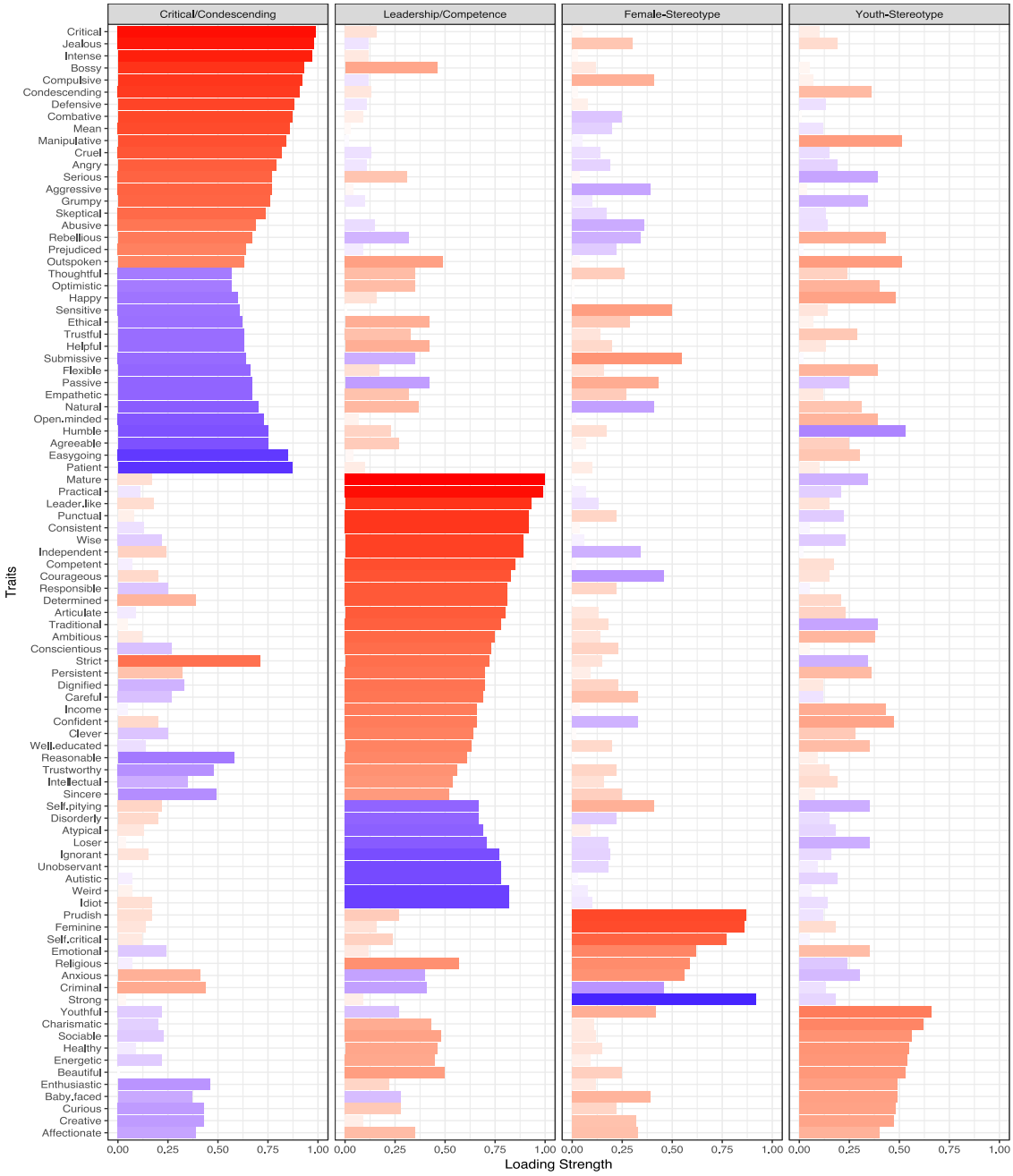


Fig. 3| Standardized factor loadings of traits. Each column plots the strength of the factor loadings (x-axis, absolute value) across all 92 traits (y-axis). The color of the bar indicates the sign of the loading (red for positive and blue for negative; more saturated for higher absolute values).

Study 2

Reliability and consensus of trait ratings

Following our preregistered data exclusion plan, criteria A to C were applied for the initial inquiry of reliability (Methods; criterion D was not applied in the current analysis because it imposed a strict lower bound on reliabilities to ensure data quality, which might lead to an overestimation of the reliability of the sample). Of the full sample with a preregistered size of $N = 30$ participants and $L = 300{,}000$ ratings at each of 7 locations ($N = 210$ total), we excluded from further analysis $n = 1$ participant in India and $l = 24{,}236$ ratings in North America, $l = 2{,}507$ ratings in Latvia, $l = 16{,}366$ ratings in Peru, $l = 3{,}178$ ratings in the Philippines, $l = 14{,}389$ ratings in India, $l = 9{,}117$ ratings in Kenya, and $l = 4{,}096$ ratings in Gaza (Methods).

All participants at all locations rated a subset of twenty traits twice across all 100 faces (Methods). Performing analyses identical to those in Study 1 on the seven datasets from the current study ($l = 100$ pairs of ratings across faces per participant for ca. $n = 28$ participants at each location after data exclusion) we found acceptable within-subject test-retest reliabilities at all locations (except for the traits *competent, religious, anxious, and critical* in India [rs = 0.18, 0.18, 0.19, 0.19] and the trait *anxious* in Peru [r = 0.19]; see Supplementary Fig. 6). As hypothesized in our preregistration, across all locations, ratings of traits that were related to physical appearance had higher within-subject test-retest reliabilities (e.g., *feminine, youthful, healthy*, with mean rs = 0.74, 0.57, 0.51, respectively) than traits that were more abstract (e.g., *critical, anxious, religious*, with mean rs =0.31, 0.32, 0.33, respectively), corroborating findings from Study 1. Interestingly, data from North America, Latvia, and Kenya revealed that ratings of *autistic* had a surprisingly high

within-subject test-retest reliability (mean $r = 0.56$ across the three datasets in the present study; $r = 0.64$ in Study 1).

Next, we inspected the between-subject consensus of trait ratings, again using analyses identical to those in Study 1. Assessment of between-subject consensus at each location used data from all participants within the same location ($l = 100$ ratings per participant for 100 faces from ca. $n = 28$ participants per trait after data exclusion in each location). Assessment of cross-cultural consensus used data from all participants across seven locations. As hypothesized in our preregistration, traits that were related to physical appearance such as *feminine*, *youthful*, *beautiful*, and *baby-faced* showed high between-subject consensus in all seven locations and high cross-cultural consensus across all locations (Supplementary Fig. 7; all ICCs > 0.86). At the other extreme, some locations had trait ratings with near-zero consensus within that location (*compulsive* in Gaza, *prudish* in India and Kenya, *self-critical* in Gaza and the Philippines). This stood in contrast to the Study 1 sample (ca. $n = 58$ participants per trait after data exclusion, who were white and located in the U.S.), who had ICCs > 0.61 for all the one hundred traits, and to the Study 2 samples from North America (ca. $n = 27$ participants per trait after data exclusion; ICCs > 0.61 for all the eighty tested traits) and Latvia (ca. $n = 28$ participants per trait after data exclusion; ICCs > 0.50 for all the eighty tested traits).

Dimensions of trait attributions across cultures

To establish the reproducibility and generalizability of the four dimensions we found in Study 1, we next carried out an EFA on the aggregate-level data at each location in Study 2 from those participants who showed acceptable test-retest reliability in their ratings. Therefore, we further applied preregistered exclusion criterion D to the seven datasets: thirty

participants across seven locations whose test-retest reliabilities for more than half of the retested traits were below 0.2 were excluded (Methods). In sum, after applying preregistered data exclusion criteria A to D, thirty-one participants across seven locations were excluded for further analysis (n = 3 for North America, n = 2 for Latvia, n = 7 for Peru , n = 3 for the Philippines, n = 10 for India, n = 2 for Kenya, and n = 4 for Gaza).

We confirmed that after this stricter data exclusion procedure, ratings of all twenty retested traits continued to have acceptable within-subject test-retest reliabilities (ranges were [0.41, 0.74] for North America, [0.41,0.85] for Latvia, [0.24, 0.78] for Peru, [0.27, 0.76] for the Philippines, [0.23, 0.78] for India, [0.26, 0.85] for Kenya, and [0.30, 0.60] for Gaza). We again verified sampling adequacy (Kaiser-Meyer-Olkin test of sampling adequacy: MSAs = 0.87 for North America, 0.83 for Latvia, 0.84 for Peru, 0.88 for the Philippines, 0.76 for India, 0.89 for Kenya, and 0.86 for Gaza) and sphericity (Bartlett's ps $< .001$).

Before performing detailed dimensional analyses, we first compared the overall correlation structures of trait ratings from the seven locations with those from Study 1 (with a subset of data from Study 1 containing the eighty traits matching those in Study 2), using representational similarity analysis (RSA) as previous research has done to compare psychological spaces across different domains[52]. The RSA showed highly similar psychological spaces between the dataset in Study 1 and all seven datasets in Study 2 (Fig. 4; essentially identical results were obtained when we used Fisher z-transformed correlations in the analysis).

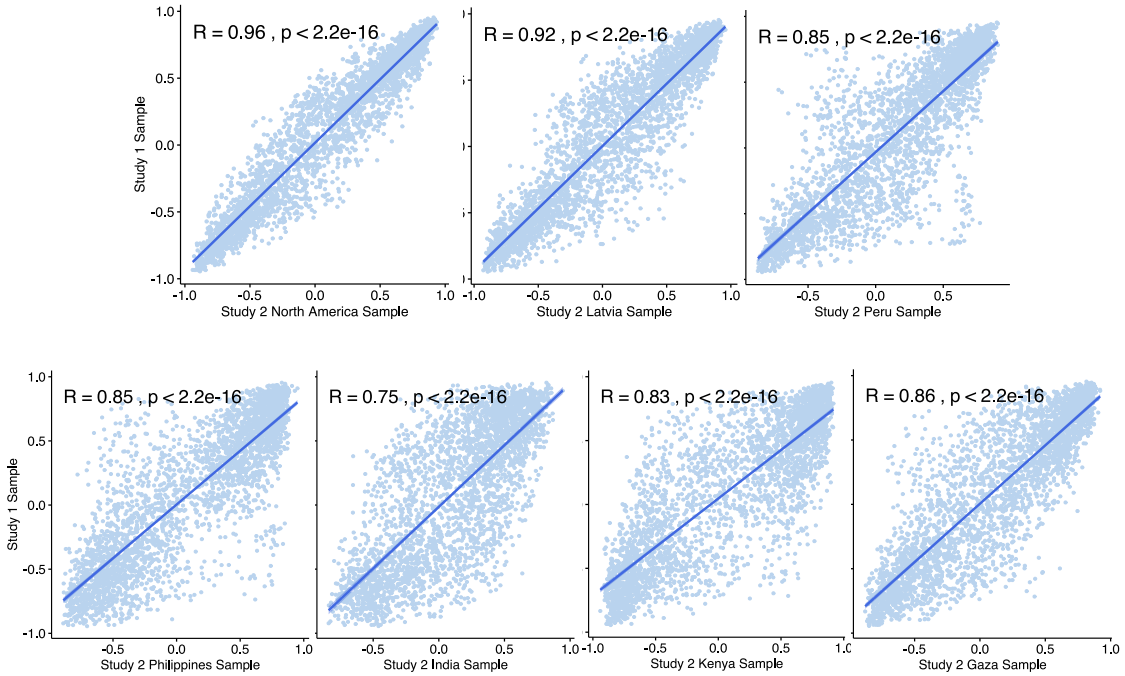


Fig. 4| Representational similarity analysis of trait ratings between the sample in Study 1 and each of the 7 samples in Study 2. Each point indicates the pairwise similarity (the Pearson correlation coefficient) of a pair of trait ratings. Lines are linear regressions of the y-axis on the x-axis. The very tight light blue area around each line indicates the 95% confidence interval. We first computed the similarity matrix for each dataset (each dataset was m by n dimensional with aggregate ratings for m = 100 faces by n = 80 traits, which produced an n by n dimensional similarity matrix). Since the traits in all datasets were measured on the same scale, similarities were assessed with Pearson correlations. We then represented each dataset with a vector consisting of all the unique pairwise similarity values in its similarity matrix (l = 3,160 similarity values per dataset). Finally, we computed the similarity between the vectors representing the seven datasets in Study 2 and the vector representing the dataset in Study 1. Fisher's z transformation was also performed to correct for vector skewness; results corroborated those shown in the figure (R = 0.95, 95% CI [0.95, 0.96] for North America; R = 0.92, 95% CI [0.91, 0.92] for Latvia; R = 0.86, 95% CI [0.85, 0.86] for Peru; R = 0.83, 95% CI [0.82, 0.84] for the Philippines; R = 0.75, 95% CI [0.73, 0.76] for Indian; R = 0.83, 95% CI [0.81, 0.84] for Kenya; R = 0.86, 95% CI [0.85, 0.87] for Gaza).

Next, we further evaluated the dimensionality of trait attributions for each of our 7 samples in Study 2. Following our preregistered data analysis plan, we used parallel analysis as in Study 1 to determine the optimal number of factors for each of the seven datasets (an m by n dimensional dataset with aggregate ratings for m = 100 faces by n = 80 traits per location). Parallel analysis, together with Cattell's scree test and the optimal coordinates index, provided evidence for a four-factor solution in five of the seven locations (North America, Latvia, Peru, the Philippines, India), whereas a three-factor solution was suggested for Kenya and Gaza (Supplementary Fig. 5c-i).

We applied exploratory factor analysis to further understand the factor structures of the seven datasets as preregistered, using the same estimation procedure (*MinRes*) and the same factor rotation method (*oblimin*) as in Study 1. An examination of the standardized loadings of each trait on the factors (Supplementary Fig. 8a-g) indicated that for four of the seven samples (North America, Latvia, Peru, and the Philippines), the four factors were critical/condescending, leadership/competence, female-stereotype, and youth-stereotype, reproducing the four dimensions found in Study 1.

For North America, these four factors each accounted for 32%, 35%, 10%, and 7% of the common variance in the data and were weakly correlated (r_{12} = -0.31, r_{13} = -0.30, r_{14} = -0.20, r_{23} = 0.14, r_{24} = 0.03, r_{34} = 0.03). For Latvia, these four factors each accounted for 24%, 41%, 9%, and 11% of the common variance in the data and were weakly correlated (r_{12} = -0.16, r_{13} = -0.22, r_{14} = -0.32, r_{23} = 0.06, r_{24} = 0.17, r_{34} = 0.03). For Peru, these four factors each accounted for 24%, 35%, 8%, and 10% of the common variance in the data and were moderately correlated (r_{12} = -0.07, r_{13} = -0.31, r_{14} = -0.39, r_{23} = 0.29, r_{24} = 0.35, r_{34} = 0.19). For the Philippines, these four factors each accounted for 27%, 37%, 9%, and 6% of

the common variance in the data, and were moderately correlated (r_{12} = -0.25, r_{13} = -0.36, r_{14} = -0.42, r_{23} = 0.32, r_{24} = 0.15, r_{34} = 0.19). Notably, the correlations among factors in these four datasets and that in Study 1 all shared the same directions. In addition, in all these five datasets, the correlation between female-stereotype and youth-stereotype (r_{34}) was one of the lowest across all pairs of correlations among the four factors.

For India, the data also revealed four factors, which each accounted for 14%, 31%, 16%, and 6% of the common variance and were moderately correlated (r_{12} = 0.06, r_{13} = -0.44, r_{14} = 0.23, r_{23} = 0.36, r_{24} = 0.06, r_{34} = -0.21). However, the interpretation of these four factors was not straightforward based merely on the visual examination of the factor loading matrix. The three factors for the Kenya sample each accounted for 33%, 30%, and 13% of the common variance in the data and were moderately correlated (r_{12} = -0.44, r_{13} = -0.53, r_{23} = 0.17). The three factors for the Gaza sample each accounted for 31%, 36% and 6% of the common variance in the data and were weakly correlated (r_{12} = -0.30, r_{13} = 0.25, r_{23} = 0.18). For the Kenya and Gaza samples, their first two factors resembled the first two dimensions we observed in the other samples (critical/condescending and leadership/competence) and their third factor resembled a mixture of the third and fourth dimensions we observed in the other samples (female-stereotype and youth-stereotype).

To further investigate how the dimensions uncovered from the seven samples in Study 2 related to the four dimensions found in Study 1, we calculated the Tucker index of factor congruence (Methods), which measures the cosine similarity between two sets of factor loadings. Since the factor analysis for all samples across Study 1 and Study 2 were conducted using the same methods and all datasets included ratings for the same set of 80 traits, we computed the factor congruence indices using the factor loadings from the

exploratory factor analyses presented above (Fig. 3 and Supplementary Fig. 8a-g; for Study 1 we used the subset of factor loadings for the 80 traits—we also confirmed that an EFA on these 80 traits alone reproduced the four dimensions found from the 92 traits, see Supplementary Fig. 8h). The factor congruence results (Fig. 5) together with the exploratory factor analyses (Supplementary Fig. 8a-g) confirmed that four of the seven samples (North America, Latvia, Peru, and the Philippines) reproduced the four dimensions of trait attributions from faces found in Study 1: critical/condescending, leadership/competence, and female-stereotype, and youth-stereotype. The three factors uncovered from the India and Kenya samples resembled the first three of the four dimensions. For the Gaza sample, its first two factors corresponded to the first two dimensions found in the other samples; and its third factor, in which only two traits (*mature*, *baby-faced*) showed highest loadings, resembled the youth-stereotype dimension.

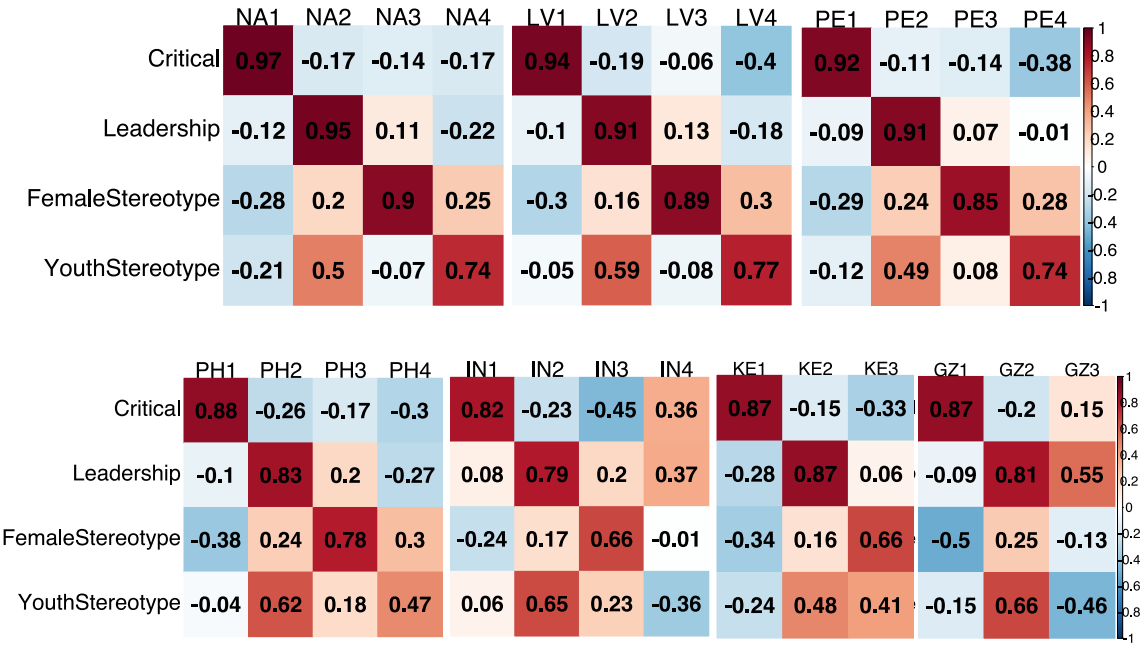


Fig. 5| Comparison of factors from Study 1 with those from Study 2. Each coefficient matrix consists of four rows, which represent the four dimensions of trait attributions found

in Study 1. The columns of each matrix represent the factors uncovered from Study 2 in the samples of North America (NA), Latvia (LV), Peru (PE), the Philippines (PH), India (IN), Kenya (KE), and Gaza (GZ), respectively. Factors within each sample were reordered (if needed) to highlight their correspondence with the four dimensions found in Study 1. The color scale shows Tucker indices of factor congruence.

Extensions

Relations between our study, the dimensions of personality, and the dimensions of face evaluation

We performed three analyses to investigate how the four dimensions we found in our studies relate to the dimensions in the literature on personality[28–36] and on face evaluation[2,38]. The personality literature suggests that personality trait attributions are represented by a five-dimensional space, and the face evaluation literature suggests that spontaneous trait attributions of unfamiliar faces are represented by a three-dimensional space. We therefore first explored what dimensions would be extracted in our own dataset if we imposed either a five-factor structure or a three-factor structure. We performed exploratory factor analysis on the data from Study 1 using the same estimation procedure and rotation method, but instead of extracting four factors, we extracted five or three. To quantify the relations among these different numbers of factors, we computed the Tucker index of factor congruence (Methods). Results (see Supplementary Table 1) showed that when a five-factor structure was imposed to explain the common variance in the dataset, the first four factors reproduced the four dimensions uncovered in Study 1 and the fifth factor also highly resembled the fourth dimension (youth-stereotype). When a three-factor structure was imposed to explain

the common variance in our dataset, these three factors reproduced the first three of our four dimensions (critical/condescending [reversed], leadership/competence, female-stereotype).

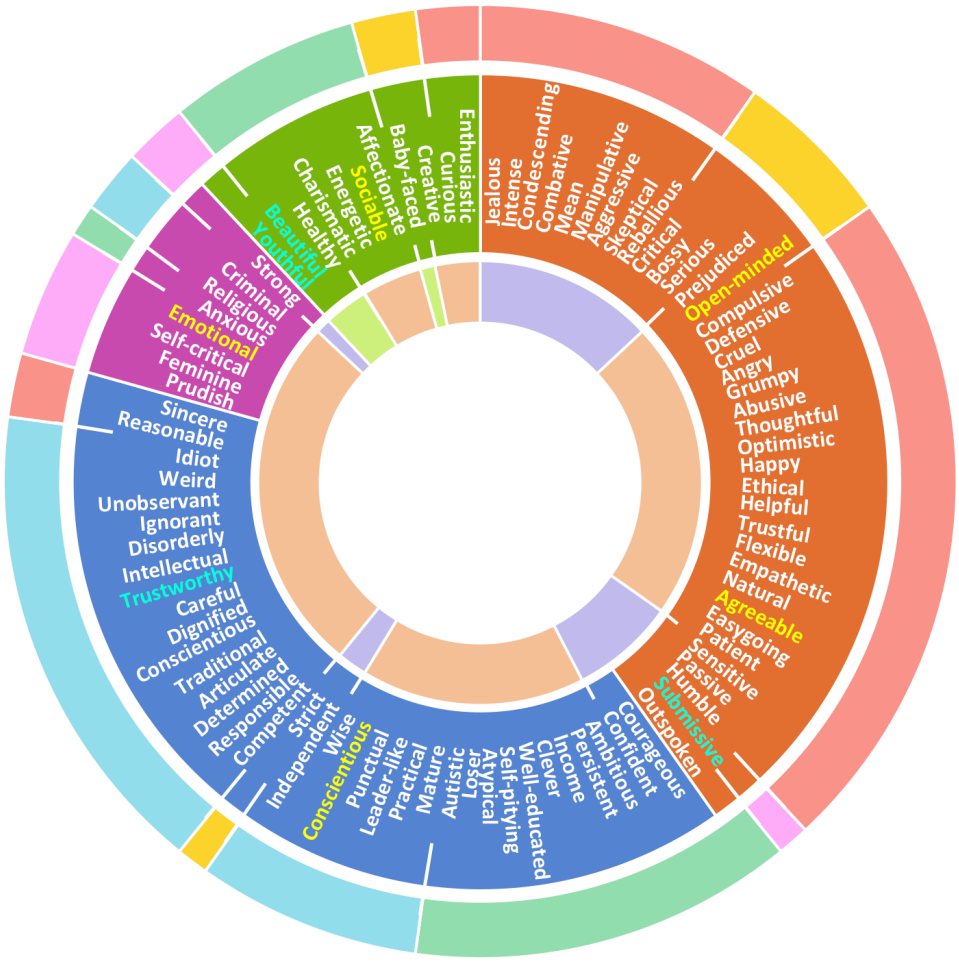
Second, we tested whether our dataset would reproduce the three face evaluation dimensions (trustworthy [or approachability, as it is referred to in some studies[38,53]], dominance, and youthful/attractiveness) discovered by previous research[38] that investigated a subset of thirteen traits. It was possible to perform such a test with our dataset because our present research followed similar experimental methods as in the literature (using faces as stimuli and collecting ratings on trait-words with Likert scales). Note that we were not able to test whether our dataset would reproduce the five personality dimensions found in the personality literature (the Big Five), since the two are not analogous (we collected attributions made about the faces of strangers, whereas personality is typically assessed with self- or peer-report questionnaires that describe one's cumulative knowledge of a person). To reproduce the three face evaluation dimensions, here we used a small subset of our Study 1 data which consisted of ratings on thirteen traits synonymous with those in the previous literature[38]. For this subset of data (with aggregate ratings for $m = 100$ faces by $n = 13$ traits), parallel analysis did not reduce dimensionality (suggesting the optimal number of factors to retain was thirteen) whereas Cattell's scree test, Kaiser's rule, and the optimal coordinates index all indicated the optimal number of factors to retain was three. We performed EFA to extract three factors following the same methods as in Study 1. An examination of the factor loading matrix indicated that these three factors indeed reproduced the three dimensions of face evaluation found in the literature[38] (Supplementary Table 2).

Third, we carried out an exploratory investigation of how the correlation structure in our comprehensive set of trait ratings would map onto the five personality dimensions and

the three face evaluation dimensions. For each of these other dimensions, we identified the one trait in our dataset whose meaning best captured one of the dimensions (we refer to these traits as "centers"). The five centers for the Big Five personality dimensions were *agreeable, conscientious, emotional, sociable,* and *open-minded* (*emotional* and *emotionally stable* shared high semantic similarity [cosine similarity = 0.55] and only the single-word adjective *emotional* was retained in our final trait set; *sociable* and *extravert* shared high semantic similarity [cosine similarity = 0.49] and only *sociable* was retained in our final trait set for its higher semantic clarity and usage frequency; *openness to experience* was replaced by a single-word adjective *open-minded*). The three centers for the three face evaluation dimensions were *trustworthy, submissive,* and *youthful/beautiful* (*submissive* and *dominant* [cosine similarity = 0.53] as well as *beautiful* and *attractive* [cosine similarity = 0.52] share high semantic similarity, and only the former were retained in our final trait set for higher semantic clarity and usage frequency).

For each of the ninety-two traits used in the exploratory factor analysis in Study 1, we calculated its distance to each center ($1 - |corr(trait, center)|$; the distance to the center *youthful/beautiful* was the averaged distance to *youthful* and *beautiful*). We then classified each trait as belonging to the center it was closest to among all centers of a given dimensional framework (five or three). These classification results (Fig. 6) showed that, compared to the Big Five personality dimensions, our critical/condescending dimension resembled the agreeableness dimension, our leadership/competence dimension resembled the conscientiousness dimension, our female-stereotype dimension resembled the emotional stability dimension, and our youth-stereotype dimension resembled the extraversion

dimension. Compared to the three dimensions found in the face evaluation literature, most of the traits were classified to the trustworthiness (approachability) dimension; some traits in our critical/condescending dimension (e.g., *aggressive*, *bossy*, *passive, submissive*), and *strong, strict, independent, confident,* and *courageous* were classified to the dominance dimension; *youthful*, *baby-faced*, *beautiful*, and *healthy* were classified to the youthful/attractiveness dimension.



Classifications of traits to the three face evaluation dimensions found in previous literature

Trustworthiness Dominance Youthful/Attractiveness

Traits with highest factor loadings in each of the four dimensions found in the present research

Critical/Condescending Leadership/Competence Female-stereotype Youth-stereotype

Classifications of traits to the five personality dimensions (the Big Five)

Agreeableness Conscientiousness Emotional Stability Extraversion Openness to Experience

Fig. 6| Classification of trait attributions in our study into the five personality dimensions and the three face evaluation dimensions. Each ring represents one dimensional framework. The 92 trait-words from our Study 1 are depicted in the middle ring, sorted by our four dimensions. The trait-words whose meaning best captured the dimensions in the other spaces were highlighted (yellow for the Big Five and green for the three face evaluation dimensions). The angular location of the trait indicates its classification to the Big Five personality dimensions (outer ring) and the three face evaluation dimensions (inner ring).

Relations between trait attribution dimensions and semantic dimensions

In our study, we not only provided a trait-word to participants for rating the faces (as is typical in other studies), but also a brief definition of the trait (Supplementary Methods), to help control for potential individual differences in interpreting the meaning of the trait. This design enabled us to explore relations between the dimensions of trait attributions from faces and the dimensions derived based merely on the semantic meanings of the trait-words. We represented each trait word in a 300-dimensional space based on the semantic features obtained from natural language processing analysis (Methods). Note that this approach yields similar vector representations for traits with similar or opposite meanings (e.g., the 300-dimensional vectors representing the three traits selfish, selfless, and altruistic are very similar and are positively correlated).

We performed exploratory factor analysis on the semantic-vector representations of the 92 traits that were included in the dimensional analysis in Study 1 (m = 300 semantic features by n = 92 traits). The Kaiser-Meyer-Olkin test of sampling adequacy (MSA = 0.97) and Bartlett's test of sphericity ($p < .001$) confirmed that this dataset was suitable for exploratory factor analysis. Parallel analysis together with optimal coordinates index

revealed that the optimal number of factors to retain was eleven. The same estimation procedure and factor rotation method as in Study 1 were applied. An examination of the factor loading matrix suggested that the eleven semantic dimensions underlying the common variance in the meanings of the 92 traits could be interpreted as physical appearance, curiosity, leadership, dominance, kindness, emotionally intense, temper, ethics, aggression, education, and communicativeness (Supplementary Fig. 9).

We then computed the Tucker index of factor congruence (Methods) between the four dimensions of trait attributions from faces and the eleven semantic dimensions. The congruence between the factors in the trait-attribution space and the trait-meaning space was low (all factor congruence indices were below 0.50), suggesting that the underlying dimensions of trait inferences from faces were not merely driven by the semantic similarity of the trait-words (Fig. 7). Interestingly, there were moderate similarities between our leadership/competence dimension and the semantic dimension of leadership, kindness, ethics, and communicativeness. There were also weak similarities between our female-stereotype dimension and the semantic dimension of emotionally intense; between our youth-stereotype dimension and the semantic dimension of physical appearance, leadership, and kindness; and between our critical/condescending dimension and the semantic dimension of kindness (reversed) and temper.

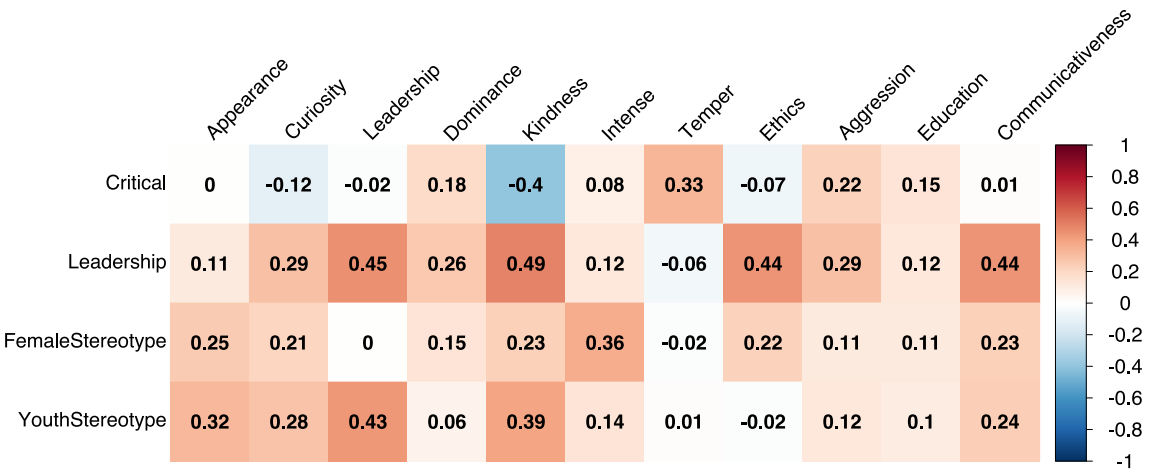


Fig. 7| Comparison between trait attribution factors and semantic factors. Each row indicates one of the four dimensions of trait attributions from faces found in Study 1. Each column indicates one of the eleven semantic dimensions extracted based merely on the meaning similarity between the traits. The color scale shows the Tucker indices of factor congruence.

Dimensions of trait attributions from individual-level data

Finally, a remaining concern was that the four dimensions uncovered so far were representative of aggregate-level data but not of any individual subject. Since we collected a complete dataset from each individual in Study 2 (an *m* by *n* dimensional matrix with individual-level ratings for $m = 100$ faces by $n = 80$ traits), we performed two analyses to compare the underlying structures between aggregate-level data and individual-level data: the representational similarity analysis [52], and a preregistered dimensional analysis (parallel analysis and EFA).

Since the number of data points for an individual participant was much less than what we used in the aggregate analyses presented so far, we needed to apply very stringent criteria to ensure the best possible data quality. Here we only included participants who had complete

trait-wise data after exclusion criteria A to D, whose ratings on all the twenty retested traits showed high within-subject test-retest reliability, who had sufficient trial-wise data, and whose data met the sampling adequacy and sphericity requirements for factor analysis (see data exclusion criteria in Methods). These stringent exclusionary criteria resulted in four final participants with the highest-quality individual-level data.

We first performed representational similarity analysis (RSA) to inspect how similar the overall correlation structures of trait attributions were between these individuals and the aggregate sample of Study 1, following the method identical to that in Study 2. The RSA showed that the overall correlation structures for the two North American participants and the Kenyan participant were highly similar to that for the sample of Study 1 (Fig. 8a). The similarity between the Latvian participant's trait attributions and those of the sample in Study 1 was lower but still significant. To address the concern that the unique pairwise similarity values were highly skewed, we also performed Fisher's Z-transformation on the similarity values before calculating the representational similarities between datasets (comparing to Study 1 sample, $R = 0.72$, 95% CI [0.70, 0.73] for North American participant #1; $R = 0.80$, 95% CI [0.78, 0.81] for North American participant #17; $R = 0.55$, 95% CI [0.52, 0.57] for Latvian participant #26; and $R = 0.72$, 95% CI [0.70, 0.73] for Kenyan participant #18).

Next, we analyzed the similarity in the dimensions of trait attributions between the individual- and aggregate-level data. We performed parallel analysis and exploratory factor analysis on each of the four individual datasets as pre-registered. Parallel analysis indicated a four-factor structure for the two North American participants and a three-factor structure for the two other participants. Exploratory factor analysis was performed to extract the

optimal number of factors as indicated by parallel analysis. The same estimation procedure and factor rotation method as in Study 1 were applied. Finally, we computed the Tucker index of factor congruence to assess the relations between the four dimensions uncovered from the aggregate-level data in Study 1 and those uncovered from the individual-level data. The individual-level data from all four participants reproduced the first three dimensions found in the aggregate-level data: critical/condescending (reversed), leadership/competence, and female-stereotype (Fig. 8b). For the two North American participants whose data revealed a four-dimensional structure, the fourth factors of both participants were moderately similar to the leadership/competence dimension and weakly similar to the youth-stereotype dimension.

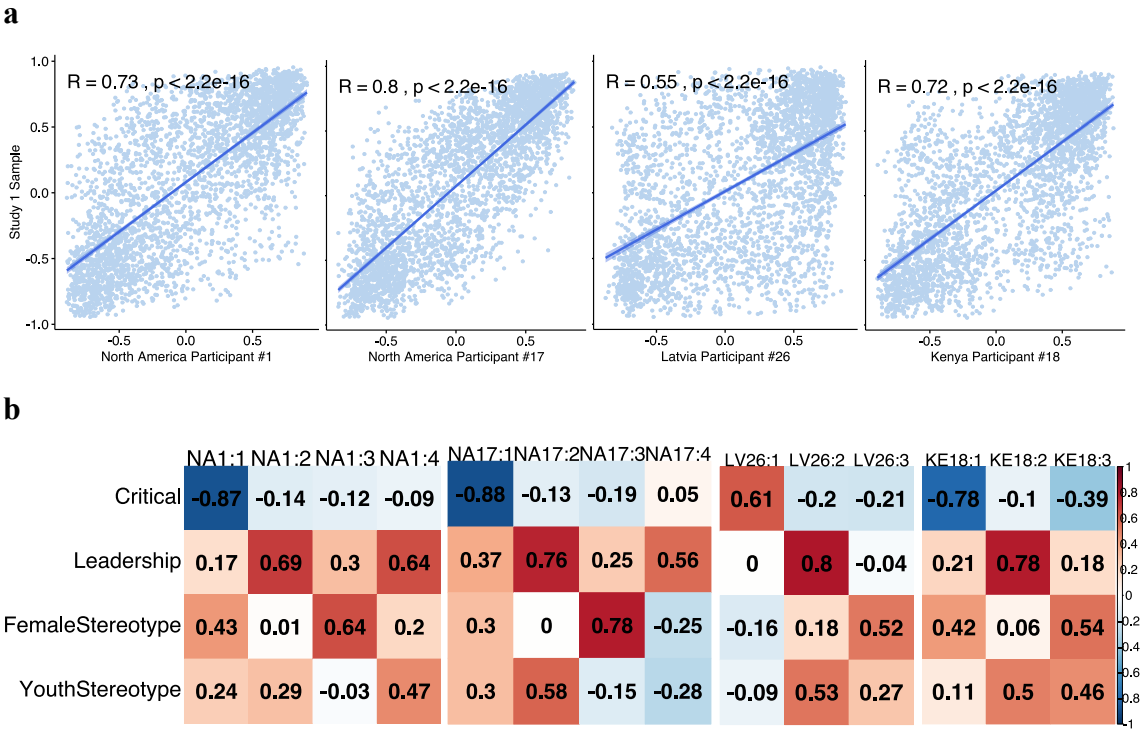


Fig. 8| Comparison of trait attributions between individual-level and aggregate-level data using representational similarity analysis (a) and dimensional analysis (b). For the scatter plots in panel (a), each point indicates the pairwise similarity (the Pearson correlation

coefficient) of a pair of trait ratings in one dataset, and lines are linear regressions of the y-axis on the x-axis, with the 95% confidence interval indicated by the light blue area around the line. For the matrices in panel (b), the rows list the four dimensions uncovered from the aggregate-level data in Study 1, and the columns in each matrix show the factors uncovered from the individual-level data of North American participant #1 (NA1), North American participant #17 (NA17), Latvian participant #26 (LV26), and Kenyan participant #18 (KE18), respectively. The color scale shows the Tucker indices of factor congruence.

4.3 Discussion

What is the psychological space in which humans represent trait attributions of other people based merely on their faces? We used deep neural networks to derive a novel set of 100 traits that representatively and non-redundantly spanned the entire domain of person perception, which included attributions of demographic characteristics, physical appearance, social evaluation, personality, emotion, and derogatory words (Fig. 1). We found that four dimensions best characterized these trait attributions from faces: critical/condescending, leadership/competence, female-stereotype, and youth-stereotype (Fig. 3). These four dimensions were reproduced (Fig. 4 and Fig. 5) across samples from different countries (U.S., Canada, Latvia, Peru, the Philippines) and with somewhat different experimental procedures (e.g., online experiments collecting sparse ratings in Study 1 versus on-site experiments collecting complete ratings per participant in Study 2, a subset of 92 traits in Study 1 versus a subset of 80 traits in Study 2, experiments administered in English in Study 1 versus Spanish in Study 2 for Peru).

Our research builds on two lines of prior work that have been largely unrelated. The investigation of individual differences in personality, mainly based on data from self-report

questionnaires, suggests that five dimensions represent individual differences in personality[25–27]. On the other hand, the study of how we make first impressions from faces has examined a set of traits (typically 13 to 25 traits)[1,38,47] that are frequently and spontaneously inferred from faces and suggests that three dimensions represent face evaluation[2,38]. It is important to note that the traits most representative of these three dimensions (i.e., the labels of the dimensions: trustworthiness/approachability, dominance, youthful/attractiveness) are all categorized as social evaluative traits, not personality traits[23,27]. Given these different dimensional frameworks, one might expect the comprehensive space for trait attributions from faces to consist of eight dimensions that combine the five personality dimensions and three social evaluation dimensions[24], or to reduce to one of the two extant spaces (five or three dimensions). Instead, we discovered a novel four-dimensional space. Notably, these four dimensions were not merely personality dimensions nor merely social evaluation dimensions. Instead, they featured an integration of personality, social evaluation, demographic characteristics, and physical appearance. This was reflected in the correspondence between our four dimensions with the Big Five personality dimensions and the three face evaluation dimensions (Fig. 6).

Our results also have implications for the literature on dimensions of social cognition (yet another dimensional framework of person perception). This literature theorizes that warmth and competence are two universal dimensions of social cognition, which arose in the evolution of social behavior. For example, when encountering a stranger, an individual needs to determine, first, the intent of the stranger (warmth), and then the ability of the stranger to execute those intentions (competence)[54]. In our research, critical/condescending (resembling the warmth dimension [reverse]) and leadership/competence (resembling the

competence dimension) were demonstrated to be two of the key dimensions of person perception from faces (Fig. 3).

Our research also dovetails with an emerging literature that aims to provide a mechanistic understanding of social attributions from faces by integrating top-down and bottom-up cognitive processes[55–57]. A number of studies argue that social categorization (e.g., sex, race) is activated not only by lower-level visual features ("bottom-up"; e.g., facial structure) but also by higher-level social knowledge ("top-down"; e.g., learned stereotypes) that shape subsequent social attributions[58] (however, whether these top-down effects penetrate visual perception *per se* is still under debate[59]). We found that female-stereotype and youth-stereotype were two key dimensions of person perception from faces (Fig. 3; regressions of trait ratings on the gender and age of the facial identities while controlling for 30 physiognomic features also confirmed that gender and age played a critical role). These two dimensions underlie the common variance in attributing a combination of physical traits (e.g., *feminine, youthful*, whose attributions were likely to rely more on low-level visual features) and personality traits (e.g., *emotional, energetic*, whose attributions were likely to rely more on higher-level social cognitive processes), which suggests that trait attributions about a person from the face require an integration between top-down and bottom-up processes.

Besides social categorization, another top-down process that potentially plays a highly influential role in how people attribute traits to others based on faces is the semantic relatedness between trait-words. It remains inconclusive in the literature[60–62] how much the correlation structure among trait attributions is driven by the visual similarities of the faces versus the semantic similarities of the trait-words. For instance, the high correlation between

attributions of *critical* and *mean* (r = 0.86) might arise because faces that are perceived to be *critical* share similar facial features with faces that are perceived to be *mean*[63] or because the semantic similarity between *critical* and *mean* nudges converging attributions from faces[64] (or both). Applying natural language processing techniques to the trait-words and their standard definitions we provided to participants, we found that the dimensions of trait-words that described their semantic correlations were not the same as our four dimensions that described the correlation structure of attributions from faces (Fig. 7), though the four dimensions were moderately correlated with a few semantic dimensions (factor congruence ranged from -0.40 to 0.49).

While we observed a high representational similarity across different participant samples in terms of the correlation structure of trait attributions (Fig. 4), and a generally similar four-dimensional space in which different samples represented these trait attributions (Fig. 5), we also noticed some variations across cultures and individuals. For example, data from the samples in India, Kenya, and Gaza only reproduced three of the four dimensions (Fig. 5). For the individual-level data in Study 2, our analyses suggested that the two individuals from North America represented these trait attributions in a four-dimensional space, while a three-dimensional space was found for the Latvian participant and the Kenyan participant (Fig. 8). These findings highlight potential cultural and individual differences that will be important to explore in greater detail in future studies.

We also assessed test-retest reliability and consensus (Fig. 2; note that these measures have no implications for the real-life accuracy of these attributions). Among the 100 traits, thirty could be highly reliably and consensually inferred from faces (rs $\geq$ 0.5 and ICCs > 0.73;

traits concerning physical appearance showed the greatest reliability and consensus) and another thirty-five traits had moderate reliability and consensus ($rs > 0.35$ and ICCs > 0.65). Given that many impactful social outcomes in the real world are generated from repeated decisions within individuals (e.g., a manager interviewing multiple job candidates and making employment decisions again and again over time) or collective decisions across individuals (e.g., a group of citizens casting their votes on a political candidate), it is plausible that traits that are more reliably and consensually inferred from faces would have greater social impacts in everyday life.

Our study has several limitations. First, we only included faces that were white, frontal, with direct gaze, with neutral facial expressions, and without any glasses or hats obscuring the face. Thus, despite our maximum variation sampling of faces, our stimuli were unrepresentative of what we often see in real life. While a large literature[17] documents the context-dependency of trait attributions from faces, and the presence of stereotypes about certain groups, our study was not aimed at investigating these factors here. Nonetheless, we corroborated evidence for a sex bias in that the interpretation of the fourth dimension was different for male faces versus female faces (even though trait attributions from male faces alone and female faces alone both revealed a four-dimensional structure). Whereas this fourth dimension for male faces was related to aggression, it was related to emotion for female faces (Supplementary Fig. 10). Future research using more diverse face stimuli with a wider range of races and ages, as well as faces in ambient photos, might uncover additional or different dimensions of person perception from faces[17].

Second, although we aimed at a most comprehensive dataset, we still used only 100 traits and (perhaps more importantly) only 100 faces. Our maximum variability sampling

aimed to span the face-structural space as uniformly as possible in selecting these stimuli, but they likely omit important physiognomic features that could influence trait attributions. Furthermore, the time required to rate the complete set precluded obtaining full datasets for each participant in Study 1, making all analyses in that study dependent on aggregate data. We addressed this limitation in Study 2, allowing us to verify the dimensions of these trait attributions at the individual-level. However, motivating participants to faithfully work on a series of experiments over multiple visits was practically very challenging. After applying strict exclusion criteria, we were left with only four individuals who had high-quality data on all the faces and all the traits. Therefore, we were only able to provide a preliminary inquiry into potential individual differences in the dimensionality of trait attributions from faces (Fig. 8). Taken together, the above limitations sum to a cautionary note: while our study aimed at the most comprehensive assessment of the dimensionality of person perception from faces, it still falls considerably short of elucidating what different people might infer from faces in the real world.

Third, we refrain from drawing any strong conclusions about cultural differences in our study. It is notoriously difficult to assure specific cultural exposure for participants, and we make no such claims here. Instead, our Study 2 was intended to extend the generalizability of our findings by providing a more culturally diverse participant set, and to collect dense individual-level data. The somewhat different factor structures we found in some countries should be considered as exploratory results that could motivate larger-scale studies focused on cultural effects in the future.

Our study makes recommendations for trait and face stimuli that could be used for future studies. Since it is practically very challenging to administer our complete set of 100

traits, a subset could be selected according to their contributions to explaining common variance (for the entire trait space or for a particular dimension, depending on the research interest; Fig. 3). Finally, we conclude with another broad future direction. A proximal explanation for the present findings must reside in the neural mechanisms that produce the ratings[65]. It will be interesting for future research to use neuroimaging to investigate whether inferences of different categories of traits engage different brain structures, and whether distinct patterns of neural activity represent each of the four dimensions of person perception from faces.

4.4 Methods

Traits.

Our goal was to sample the most comprehensive list of trait-words that could be used to describe people from their faces. We derived a final set of 100 traits through a series of combinations and filters. We began with an initial set of 482 adjectives from two sources. Source #1 were 435 personality adjectives from previous lexical studies of personality[25]. Source #2 were an additional 47 traits from prior face evaluation studies[1,2,47].

Many of the traits from Source #1 and #2 had similar or opposite meanings. To avoid redundancy while conserving semantic variability, we sampled 100 traits from these 482 traits based on three criteria: a trait's semantic similarity to other traits, clarity in meaning, and frequency in usage. For traits with similar meanings, clarity was the second selection criteria (the one with the highest clarity was retained). For traits with similar meanings and the same clarity, usage frequency was the third selection criteria (the one with the highest usage frequency was retained).

To quantify the semantic similarity between traits, we represented each trait with a vector of 300 computer-extracted semantic features (for word embeddings and text classification) using a state-of-the-art neural network provided within the FastText library[48] that had been trained on Common Crawl data of 600 billion words to predict the identity of a word given a context. We then applied hierarchical agglomerative clustering (HAC) on the word vectors of our traits based on their cosine distances, and used visual inspection of the dendrogram produced by HAC to assess semantic similarity. Subsequent to this step, we quantified the clarity of word meanings by obtaining ratings of clarity from an independent set of participants tested via MTurk ($N = 31$, 17 males, Age ($M = 36$, SD = 10)). To quantify the usage frequency of a trait, we obtained the average monthly google search frequency for the bigram of each trait (i.e., the trait-word with the word *person* added after it) using the keyword research tool Keyword Everywhere.

Based on the three filters, the 482 traits integrated from Source #1 and #2 were reduced to 94 traits. We further supplemented the final trait set from two additional sources: Source #3 were 4 descriptors related to demographic and health characteristics (education, income, sexuality, autism); Source #4 were 2 frequently used derogatory words (idiot, loser). The final 100 traits used in the present research can be accessed at the Open Science Framework (https://osf.io/6p542/?view_only=fff024253b604edb832a9824cbdafa75).

Face Stimuli.

Our goal was to derive a final set of 100 face images of excellent quality drawn from multiple databases so as to best span all dimensions of structural variability. Our initial set of 909 high-resolution photographs of male and female faces were combined from three publicly available face databases (the Oslo Face Database[44], the Chicago Face Database[43],

and the Face Research Lab London Set[42]). We restricted ourselves to faces that were front-facing, with neutral facial expression and direct-gaze, and without glasses or other adornments, because facial images with these restrictions are the most common type of stimulus used. We furthermore restricted ourselves to photographs of Caucasian adults, because we were not interested in investigating race variables in this study. This yielded a set of 426 faces from the three databases.

To further reduce the number of faces while retaining maximal variability in facial structure, we sampled one hundred faces (50 females, 50 males) from this set of 426 faces using maximum variation sampling. For each image, the face region was first detected and cropped using the dlib library[45]. We then vectorized each face region with 128 computer-extracted facial features (for recognizing facial identities) using a state-of-the-art neural network provided within the dlib library that had been trained to identify individuals across millions of faces (of all different aspects and races) with very high accuracy[45]. Next, we sampled 50 female faces and 50 male faces that respectively maximized the sum of the Euclidean distances between their face vectors. Specifically, a face image was first randomly selected from the female or male sampling set, and then other images of the same gender were selected so that each new selected image had the farthest Euclidean distance from the previously selected images. We repeated this procedure with 10,000 different initializations and selected the sample with the maximum sum of Euclidean distances. We repeated the whole sampling procedure 50 times to ensure convergence of the final sample.

All 100 final faces were frontal, clear, with neutral expression, and presented at the center of the images with the eyes at the same height across the images. All photos included faces, neck, and hair, were colored, had a standard grey background, and were cropped to a

standard size and shape. The final one hundred faces can be accessed at the Open Science Framework (https://osf.io/4mvyt/?view_only=e998f9c39b6f4dcb82d15035cefd65ca).

Participants (Study 1).

We predetermined our sample size to be 60 participants per trait based on a recent study that investigated the point of stability for impression formation from faces[41]. That study analyzed a dataset containing 698,829 ratings from 6,593 participants for 3,353 facial stimuli and 24 traits[47]. The study found that for ratings assessed on a 7-point Likert scale and for a point of stability measured according to an acceptable corridor of stability of +/- 0.5 with a confidence level of 95%, a stable average rating for each of the 24 traits could be obtained in a sample with a size ranging from 18 to 42 participants. Based on these findings, we preregistered our sample size to be 60 participants for each trait (see preregistration form at https://osf.io/6p542/?view_only=fff024253b604edb832a9824cbdafa75).

Participants were recruited via Amazon Mechanical Turk (N = 1,500 (800 males), Age(M = 38 years, SD = 11), median of educational attainment was "some post-high-school, no bachelor's degree"). All participants were required to be white, native English speakers, located in the U.S., and 18 years old or older, with normal or corrected-to-normal vision, an educational attainment of high school or above, and a good MTurk participation history [approval rate ≥ 95%]). We also collected data about whether our participants were currently being treated for psychiatric or neurological illness. The majority of our participants (79.7%) were not currently being treated for any psychiatric or neurological illness. The rest were currently being treated for depression (9.8%), bipolar disorder (1.3%), anxiety or panic disorder (11.2%), obsessive compulsive disorder (0.9%), post-traumatic stress disorder (1.3%), autism spectrum disorder (0.3%), learning disability (0.1%), attention deficit

(0.9%), alcohol or drug addiction (1.0%), personality disorder (0.5%), dissociative disorder (0.1%), epilepsy (0.2%), and brain injury (0.1%). All dimensional analyses in Study 1 were repeated excluding participants who were currently being treated for any psychiatric or neurological illness, and the results corroborated those we reported in the Results—specifically, for this subset of data, the same eight traits were found to have low factorability and therefore removed from subsequent analyses; parallel analysis together with Cattell's scree test and optimal coordinates index indicated that the optimal number of factors for the remaining data was four; EFA extracting four factors from these remaining data showed that the four factors were essentially identical to the four dimensions reported in Study 1 (Tucker indices of factor congruence = 1.00, 1.00, 0.99, 0.99).

Participants (Study 2).

We preregistered to recruit participants through Digital Divide Data, a social enterprise that delivers research services, in seven countries/regions of the word: North America (U.S. and Canada), Latvia, Peru, the Philippines, India, Kenya, and Gaza. All participants were required to be between 18-40 years old, proficient in English (except participants in Peru), have been educated and completed at minimum high school, have been trained in basic computer skills, and have never visited or lived in western-culture countries (except participants in North America and Latvia). In addition, we aimed to have a roughly equal sex ratio of participants in all locations. The sample size for each location was predetermined to be 30 participants. This sample size was determined based on two criteria: first, the sample size should be big enough to ensure stable average ratings of trait inferences (for a corridor of stability of +/- 1.00 and a level of confidence of 95%, the point of stability ranged from 5 to 11 participants for the inferences of 24 traits from faces[41]); second, the sample

size should be feasible for all the seven locations given the requirements mentioned above and the availability of participants for paying multiple visits to the local offices to complete all the experiments over a 10-day period. Our preregistration can be accessed at OSF (https://osf.io/qxgmw/?view_only=fd43b2e8b25248f7b7de51b9aeae1894).

As planned, 30 individuals (15 females and 15 males) in each of the seven locations participated in our study (Age (M = 26, SD = 4) for North America; Age (M = 22, SD = 3) for Peru; Age (M = 28, SD = 5) for Latvia; Age (M = 26, SD = 5) for Gaza; Age (M = 24, SD = 2) for Kenya; Age (M = 27, SD = 6) for India; Age (M = 25, SD = 4) for Philippines). All participants were confirmed to meet the requirements mentioned above.

Procedures (Study 1). All experiments were completed online via MTurk. Considering the amount of time it would take for a participant to complete ratings for all one hundred traits and one hundred faces, we divided the experiment into 25 modules (the 100 traits were randomly shuffled once and divided into 25 modules, each consisted of 4 traits). Each participant completed one module.

To encourage participants to use the full range of the rating scale to evaluate the faces, participants were shown all the one hundred faces briefly at the beginning of a module, so that they had a sense of the range of the faces they were going to rate. In each module, participants rated all faces for each of the four traits (in random order) in the first four blocks, and then in the last (fifth) block they rerated all faces for the trait they were assigned in the first block again, thus providing sparse test-retest data for our traits. At the beginning of each block, participants were instructed on the trait they were asked to evaluate and were provided with a clear one-sentence definition of the trait (see Supplementary Method). Participants viewed the faces one by one in random order and rated each face for a trait on

a 7-point Likert scale. Each face appeared for one second. Participants could enter their ratings as soon as the photo appeared or within four seconds after the photo disappeared. Participants entered their ratings by pressing the number keys on the computer keyboard. The orientation of the Likert scale in each block was randomized across participants. At the end of the experiment, participants completed a short survey questionnaire on demographic information. A sample of our experiment instruction can be accessed at OSF (https://osf.io/6p542/?view_only=fff024253b604edb832a9824cbdafa75).

Procedures (Study 2).

All experiments were completed onsite in the Digital Divide Data local offices. Participants in North America, Latvia, Gaza, Kenya, India, and the Philippines completed the experiments in English. Participants in Peru completed all experiments in Spanish. An exact translation of the experiment instructions, trait words, and definitions of the traits from English to Spanish was provided by the professionals in the Peru office of Digital Divide Data.

Eighty of the 100 traits were used in Study 2—twenty traits were excluded for their low correlations with other traits as found in Study 1 (*sarcastic, white, thrifty, shallow, homosexual, nosey, conservative,* and *reserved*), their ambiguity or similarity in meaning as found in feedback from Study 1 (*trustful, natural, passive, reasonable, strict, enthusiastic, affectionate,* and *sincere*), and potentially inappropriate in some cultures (*idiot, loser, criminal,* and *abusive*).

Participants in all the seven countries/regions followed the same experimental procedures. Each participant provided evaluations for all the 100 faces on all the 80 traits, of which 20 traits were rated twice by each participant for test-retest reliability. The 80 traits

were divided into 20 modules, each consisting of 4 distinct traits (the 20 rerated traits were first assigned to separate modules, and then the other 60 traits were randomly assigned to the 20 modules with the constraints that the 4 traits in the same module should be balanced in valence). All participants completed all the 20 modules during multiple visits to the local offices in ten business days. Each module consisted of 5 blocks, in which participants rated one retested trait in the first and last blocks and rated the other three traits in the other blocks in random order. The experimental procedure within each module was identical to that in Study 1.

Both the English and Spanish versions of the experiment instructions, the list of the 80 traits and 20 retested traits, and the definitions of the traits can be accessed at our preregistration (https://osf.io/qxgmw/?view_only=fd43b2e8b25248f7b7de51b9aeae1894).

Data exclusion criteria (Study 1).

Data were excluded following three preregistered criteria: a. Trial-wise deletion would be done if a response was missing or timed out, or if RT was less than 100ms; b. Participant-wise deletion would be done if a participant had more than 10% of invalid trials in any block as per (a); c. Block-wise (trait-wise per participant) deletion would be done if all trials in a given block had the same rating. Our preregistration can be accessed at Open Science Framework (https://osf.io/6p542/?view_only=fff024253b604edb832a9824cbdafa75).

Data exclusion criteria (Study 2).

Two exclusion criteria were planned in the initial preregistration: a. Trial-wise deletion would be done if a response was missing or timed out, or if RT was less than 100ms; b. Block-wise (trait-wise per participant) deletion would be done if all trails in a given block had the same rating. To ensure high quality and complete data from individuals, we further

registered four exclusion criteria while data collection was underway and data had not yet been analyzed. A. Trial-wise deletion would be done if a rating was missing or timed out, or if RT was less than 400ms; B. Block-wise (trait-wise per participant) deletion would be done if (B1) a block had more than 10% ratings were missing or with RTs less than 100ms, or (B2) a block had more than 20% ratings with RTs less than 400ms, or (B3) a block had the same rating for all faces; C. Participant-wise deletion would be done if a participant's test-retest reliability for more than 25% of the retested traits were more than three standard deviations below the mean test-retest reliability of the traits as found in the sample of Study 1; D. Participant-wise deletion would be done if a participant's test-retest reliability for more than 50% of the retested traits were below 0.20. Preregistrations can be accessed at OSF (https://osf.io/qxgmw/?view_only=fd43b2e8b25248f7b7de51b9aeae1894 and https://osf.io/tbmsy/?view_only=6d8b94575bf0469fb157c89eb9292371).

For the dimensional analysis based on individual-level data (Extensions), we first applied criteria A to D and then participants who did not have complete trait-wise data (whose data has been trait-wise deleted) were dropped (e.g., a participant's ratings for a trait would be excluded entirely if there were missing responses for more than 10% of the faces according to the trait-wise exclusion criterion B). For each of the remaining participants who had complete trait-wise data, we calculated his/her test-retest reliability for each of the twenty retested traits. Participants whose test-retest reliabilities for all the twenty retested traits were greater than 0.20 were retained (n = 6 for North America, n = 1 for Peru, n = 7 for Latvia, n = 2 for Gaza, n = 4 for Kenya, n = 1 for India, and n = 0 for Philippines). For the remaining participants, we checked their trial-wise data availability: participants with complete observations (no missing data across the eighty traits) for no fewer than 80% of

the faces were retained (n = 5 for North America, n = 0 for Peru, n = 3 for Latvia, n = 0 for Gaza, n = 1 for Kenya, n = 1 for India). We further performed the Kaiser-Meyer-Olkin test of sampling adequacy and the Bartlett's test of sphericity on each remaining individual's dataset. The remaining datasets of four individuals (n = 2 for North America, n = 1 for Latvia, and n =1 for Kenya) met the requirements of factor analysis (MSA > 0.60 and Bartlett's $p < 0.05$).

Linear mixed-effect modeling.

Linear mixed-effect models are an extension of linear models that analyze non-independent data (e.g., repeated measures from an individual) by allowing for both fixed and random effects. Formally, if we denote a parameter a researcher is interested in as β, a fixed effect is a parameter that does not vary in the population and we can get an estimate of it, $\hat{\beta}$; a random effect is a parameter that is itself random, for example, according to a normal distribution with mean μ and standard deviation σ, $\beta \sim \mathcal{N}(\mu, \sigma)$. A linear mixed model can be expressed as

$$y = X\alpha + Z\beta + \varepsilon$$

where y is the outcome variable, X is the data matrix of predictor variables for estimating the fixed effects α, Z is the design matrix of unobserved variables which acts as the random complement to the fixed X, β is a vector of the random effects which acts as the random complement to the fixed α, and ε is a vector of residuals that captures the part of y that is not explained by $X\alpha + Z\beta$.

We used linear mixed-effect modeling to estimate test-retest reliability of trait attributions: we regressed individual trait ratings collected at the second time on those

collected at the first time while including individual subjects as random effects. We also used linear mixed-effect modeling to estimate the effect of four participant characteristics (gender, age, education, and psychiatric/neurological illness) on trait attributions: we regressed individual trait ratings on the dummy variable that indicated which of the median-split characteristic group the individual belonged to while including individual subjects as random effects.

Intraclass Correlation Coefficient.

The intraclass correlation coefficient (ICC) measures the reliability of ratings given by a set of *J* participants for a set of *I* stimuli. Here we used ICCs to quantify the consensus amongst participants—that is, the reliability of ratings that they produced across participants. There are two important decisions to make for computing ICC: model and type[66]. Our experimental design and measurement goal justified a two-way random-effects multiple-participants ICC (ICC(2,k)). Formally, if we denote a rating for image *i* from participant *j* as $R_{ij,}$ and represent it as an additive function of three components:

$$R_{ij} = \alpha + \beta_i + \varepsilon_{ij}$$

where α is the grand mean of the ratings, β is the effect due to the image, and ε is random noise, then ICC(2,k) can be expressed as

$$ICC = \frac{MSB - MSE}{MSB + \frac{(MSJ - MSE)}{J}}$$

where *MSB* = mean square for rows, *MSE* = mean square for error, *MSJ* = mean square for columns.

Exploratory factor analysis.

We used exploratory factor analysis (EFA) to derive the factor structure of our ratings. EFA is a commonly used statistical method for analyzing the latent construct of a relatively large set of observed variables when there is no prior hypothesis. It analyzes how each measured variable correlates with all other measured variables and extracts a relatively small number of factors to represent the common variance in the measured variables. Exploratory factor analysis is a more appropriate method in our case than other dimensional reduction methods such as principal component analysis. Formally, if we hypothesize that a set of observed variables (variables refer to traits in our case) $x_1, x_2, \dots, x_L$ arise as a set of linear combinations of K unobserved, latent, common factors $\xi_1, \xi_2, \dots, \xi_K$ where $K \ll L$, then two variables x_i and x_j sharing a common factor ξ_k can be expressed as

$$x_i = \lambda_i \xi_k + z_i$$

$$x_j = \lambda_j \xi_k + z_j$$

where z denotes the unique factor that is associated with each variable. The common and unique factors are assumed to be uncorrelated and all the unique factors are assumed to be uncorrelated and centered. In matrix terms, the model can be expressed as

$$\mathbf{x} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{z}.$$

Since the common factors are unobserved variables, the above model is not testable. However, a particular form of the variance-covariance matrix $\boldsymbol{\Sigma}$ of the observed variables implies from the above model is testable, which can be expressed as

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\Psi}$$

where $\boldsymbol{\Lambda}$ is a $L \times K$ dimensional factor loading matrix, $\boldsymbol{\Phi}$ is a $K \times K$ dimensional factor correlation matrix, and $\boldsymbol{\Psi}$ is a $L \times L$ dimensional diagonal matrix of the unique variances of

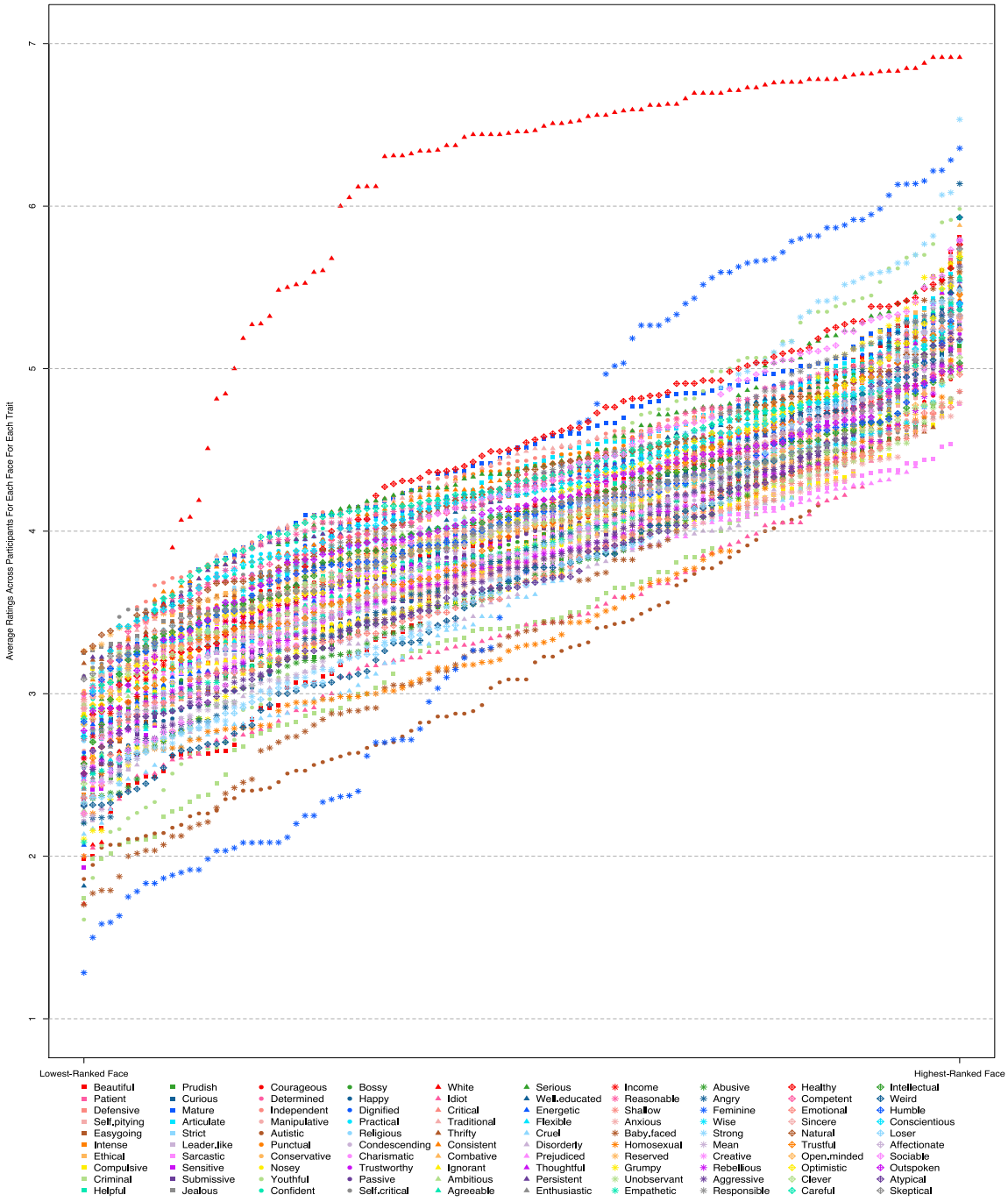
the observed variables. Given the observed variance-covariance matrix of the observed variables $\boldsymbol{S}_{L\times L}$, the goal of exploratory factor analysis is to estimate $\widehat{\boldsymbol{\Lambda}}, \widehat{\boldsymbol{\Phi}}$, and $\widehat{\boldsymbol{\Psi}}$ to approximate $\boldsymbol{\Sigma}$.

Tucker index of factor congruence.

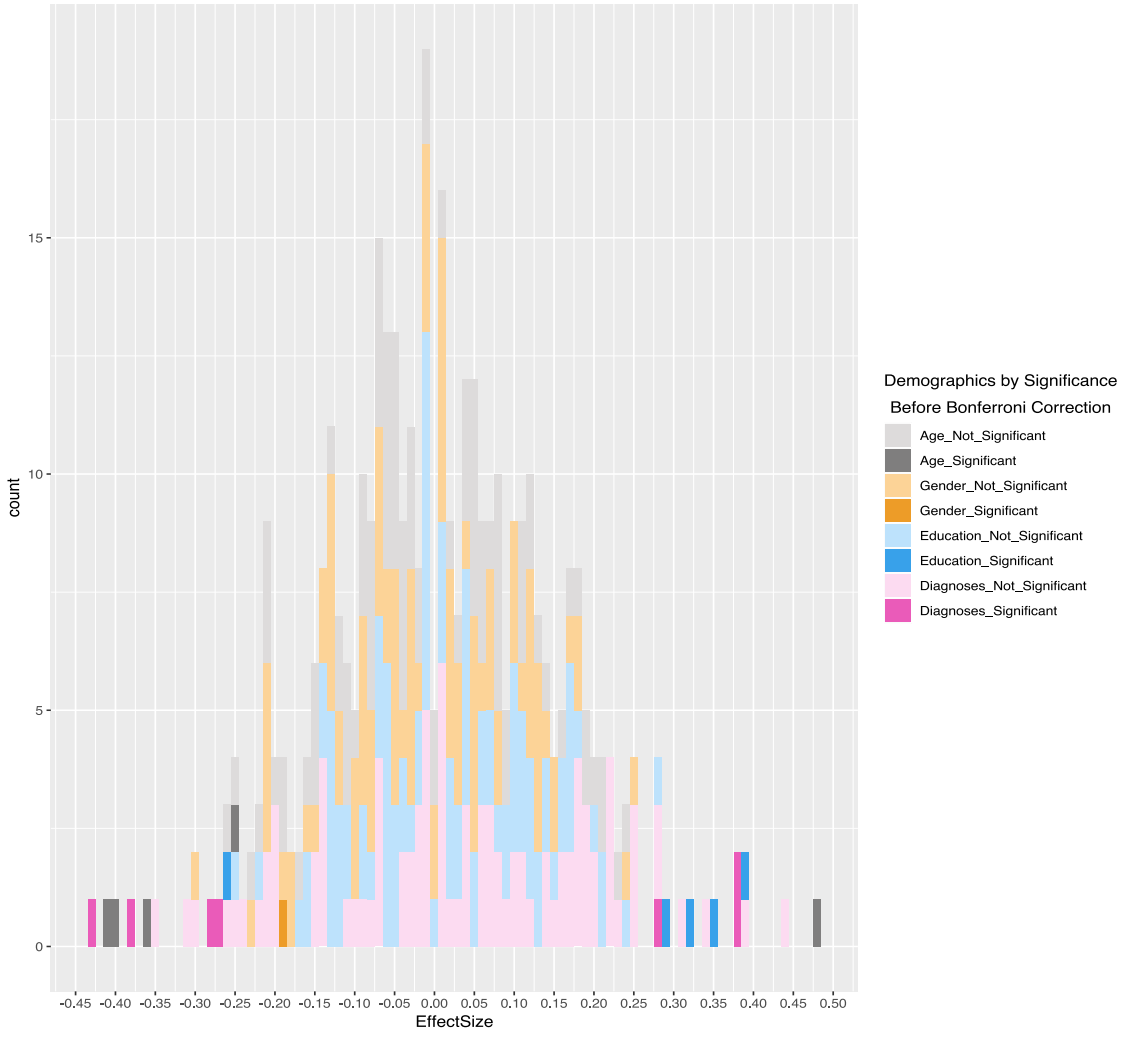
The Tucker index of factor congruence[67] assesses the similarity between factors that have been derived from a factor analysis. Formally, if we denote the loadings of two factors as F_1 and F_2, then the congruence coefficient r can be expressed as

$$r = \frac{\sum F_1 F_2}{\sqrt{\sum {F_1}^2 \sum {F_2}^2}}.$$

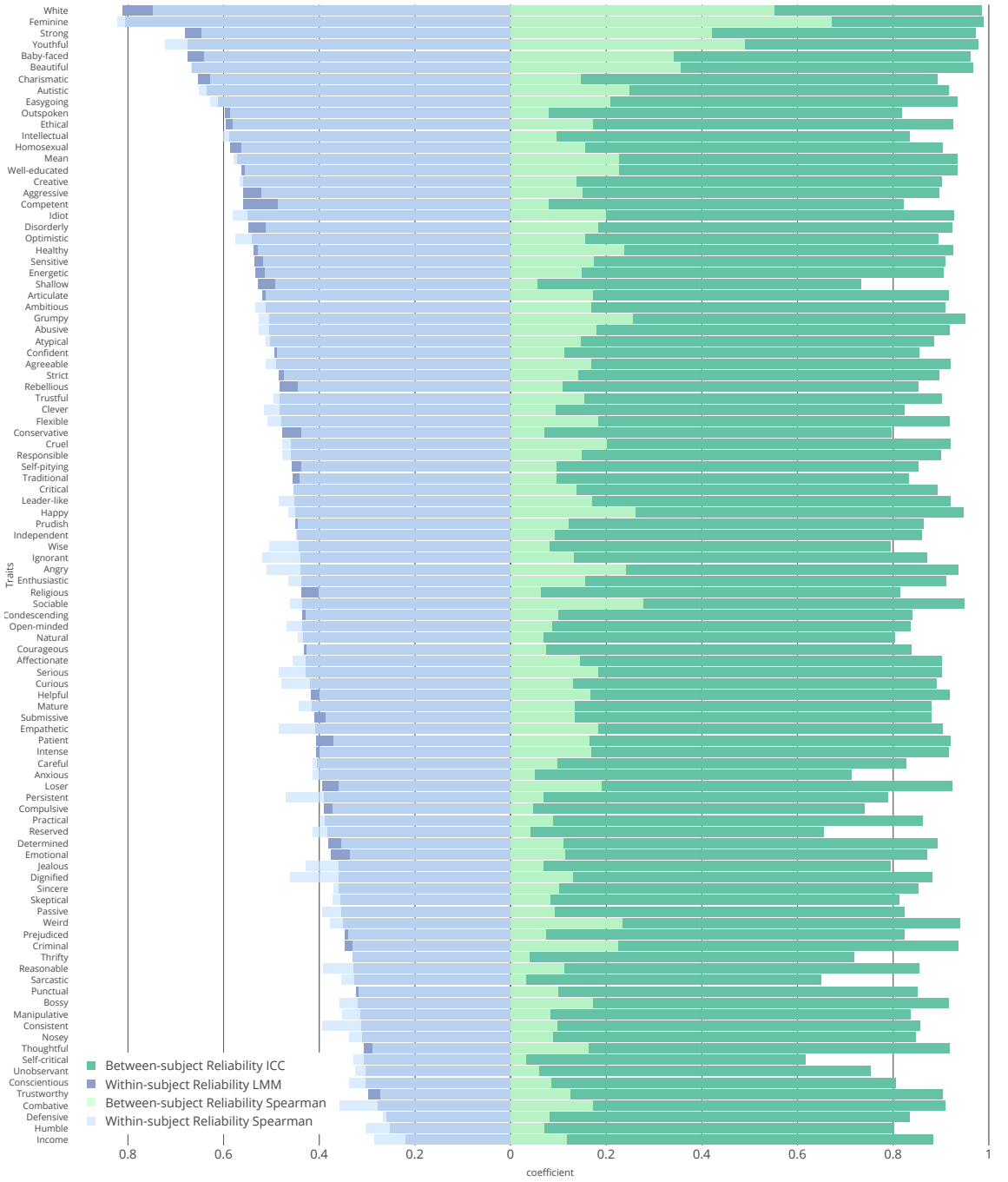
4.5 Supplementary Information



Supplementary Figure 1: Distributions of average ratings per face for the 100 traits. Each point plots the mean rating averaged across all participants for one face on a trait. Average ratings for each trait are indicated by points with a unique combination of color and shape. Average ratings are sorted from low to high across faces per trait (from the face with the lowest average rating to the face with the highest average rating).



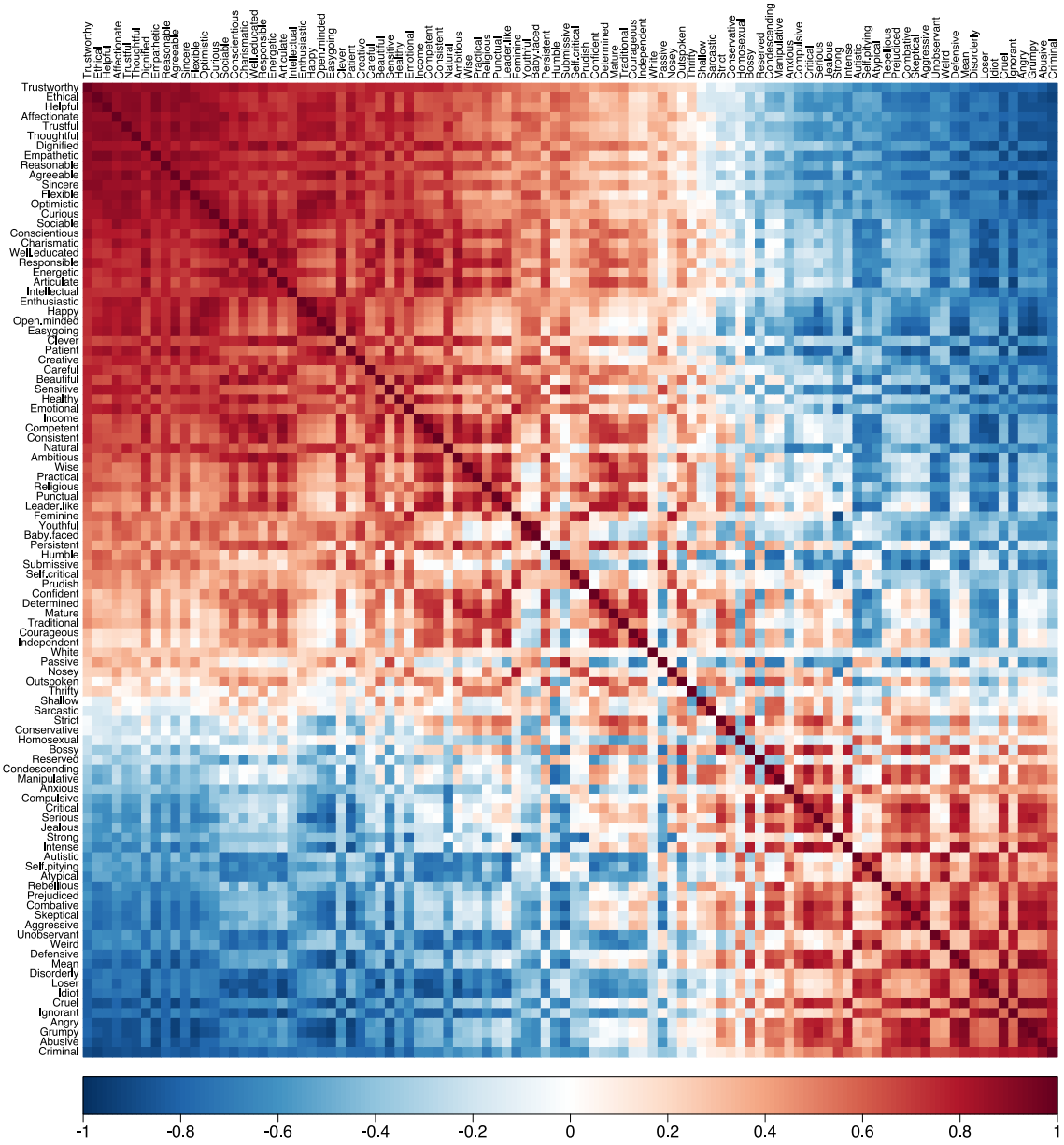
Supplementary Figure 2: Distribution of effect sizes for four characteristics of participants. For each of the four characteristics, participants were median-split into two groups (i.e., age {low, high}, gender {male, female}, education {low, high}, diagnoses {yes, no}). The effects were estimated with linear mixed-effect modeling for each characteristic and each trait respectively (n = 400 mixed-effect models in total), where the effect of the characteristic was treated as a fixed effect and the effects of individual participants were treated as random effects. The colors of the bars indicate the characteristics; low saturated colors indicate non-significant effects and high saturated colors indicate significant effects before Bonferroni correction. There were 18 significant effects before Bonferroni correction, which were found for the following specific trait attributions (sorted from the most negative to the most positive effect sizes per characteristic): participants' age on attributing *natural, strong, prudish, wise,* and *healthy*; participants' gender on attributing *shallow*; participants' education on attributing *rebellious, trustful, nosey, open-minded,* and *natural*; participants' psychiatric/neurological illness on attributing *articulate, intellectual, feminine, compulsive, youthful, reserved,* and *conservative*. After Bonferroni correction, none of the four characteristics showed a significant effect on any trait attribution from faces.



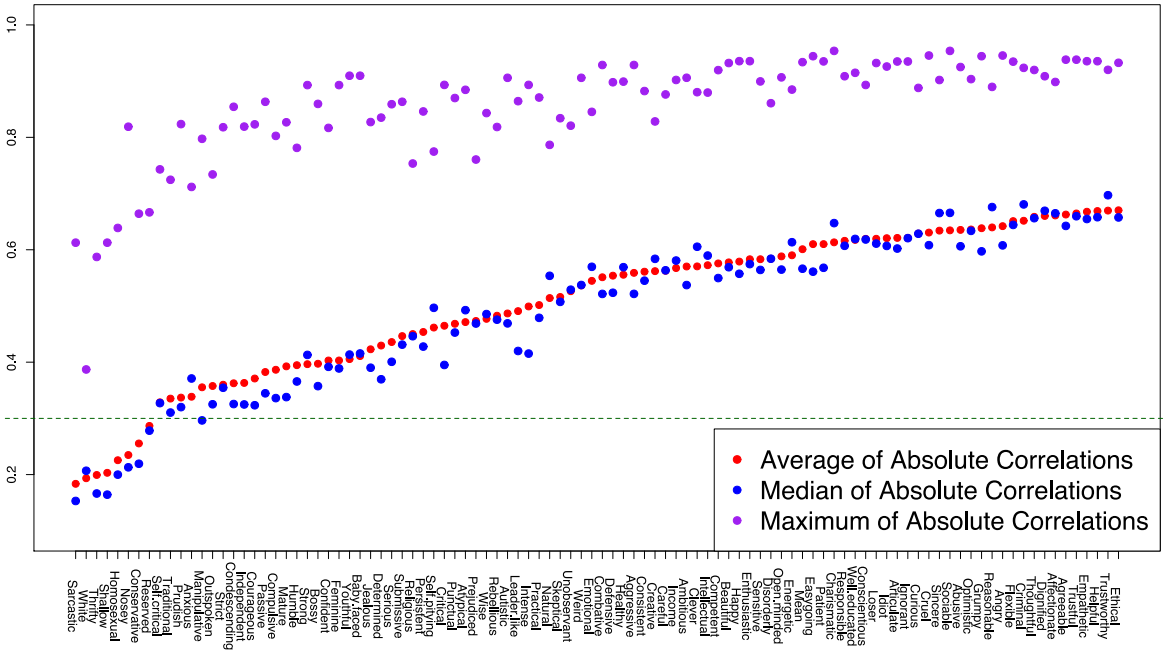
Supplementary Figure 3: Within-subject test-retest reliability and between-subject consensus of attributions across 100 traits. Each row plots the within-subject test-retest reliability (blue) and the between-subject consensus (green) for one trait. The length of the bar indicates the magnitude of the coefficient. The two different saturations of each color indicate the two different methods for computing the coefficients: high saturated colors for

preregistered methods and low saturated colors for Spearman correlations. Spearman correlations were used to reassess both quantities (beyond what was preregistered) in order to compare within-subject test-retest reliability and between-subject consensus on a common metric. For the between-subject consensus, we calculated for each pair of participants the Spearman correlation between their ratings across all faces for the same trait (ca. l = 100 pairs of ratings); these correlations were converted to z-scores using Fisher's z-transformation and then averaged across all pairs of participants (ca. n = 1653 pairs of participants per trait after data exclusion); the averaged z-scores were converted back using Fisher's z-transformation to obtain the averaged Spearman correlation per trait. For the within-subject reliability, we calculated for each participant the Spearman correlation between his/her repeated ratings across all faces for the same trait (ca. l = 100 pairs of ratings per participant), and averaged those correlations (with Fisher's z-transformation) across all participants (ca. n = 15 participants per trait after data exclusion).

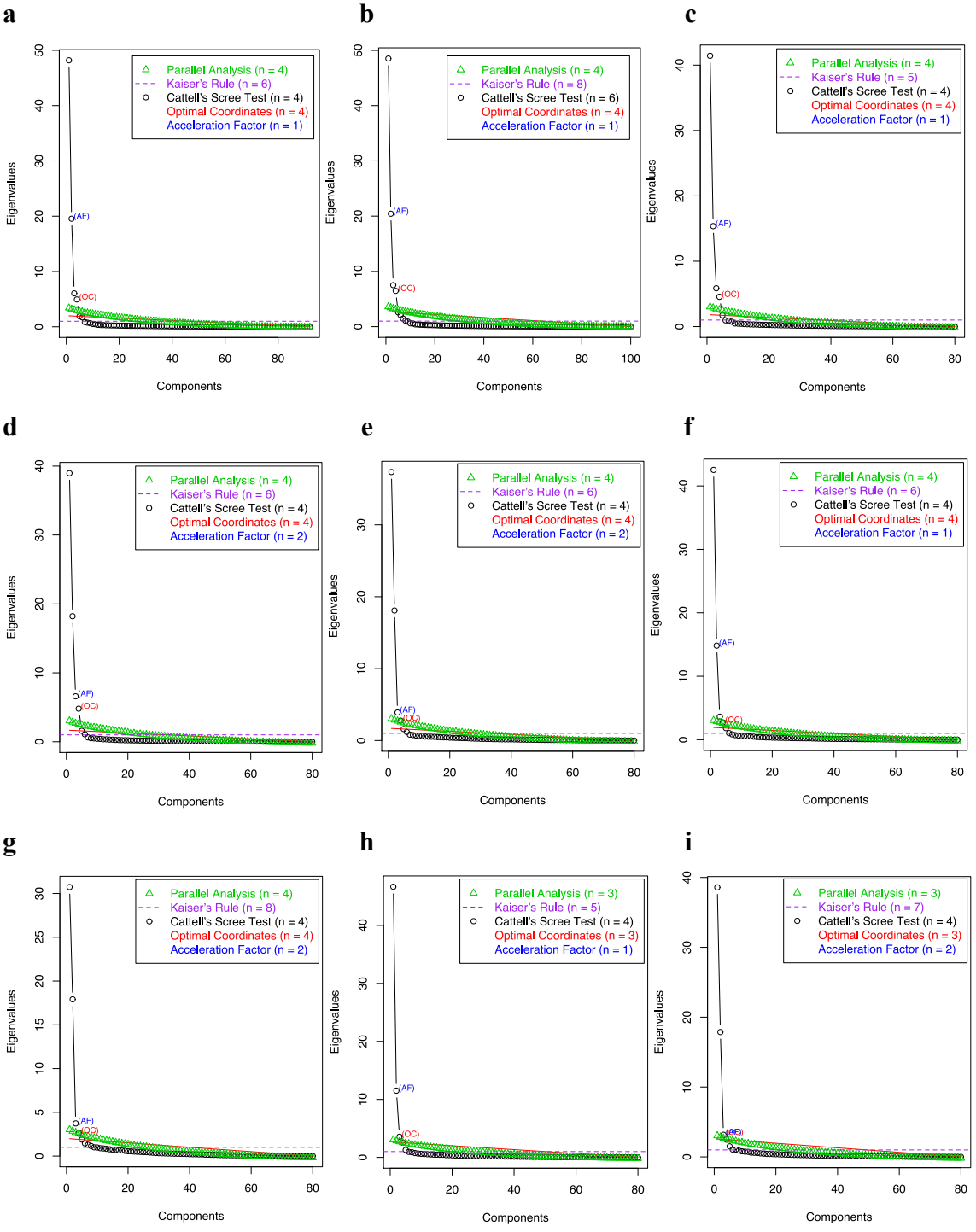
a


Trustworthy
Ethical
Helpful
Affectionate
Trustful
Thoughtful
Dignified
Empathetic
Reasonable
Agreeable
Sincere
Flexible
Optimistic
Curious
Sociable
Conscientious
Charismatic
Well.educated
Responsible
Energetic
Articulate
Intellectual
Enthusiastic
Happy
Open.minded
Easygoing
Clever
Patient
Creative
Careful
Beautiful
Sensitive
Healthy
Emotional
Income
Competent
Consistent
Natural
Ambitious
Wise
Practical
Religious
Punctual
Leader.like
Feminine
Youthful
Baby.faced
Persistent
Humble
Submissive
Self.critical
Prudish
Confident
Determined
Mature
Traditional
Courageous
Independent
White
Passive
Nosey
Outspoken
Thrifty
Shallow
Sarcastic
Strict
Conservative
Homosexual
Bossy
Reserved
Condescending
Manipulative
Anxious
Compulsive
Critical
Serious
Jealous
Strong
Intense
Autistic
Self.pitying
Atypical
Rebellious
Prejudiced
Combative
Skeptical
Aggressive
Unobservant
Weird
Defensive
Mean
Disorderly
Loser
Idiot
Cruel
Ignorant
Angry
Grumpy
Abusive
Criminal
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

b

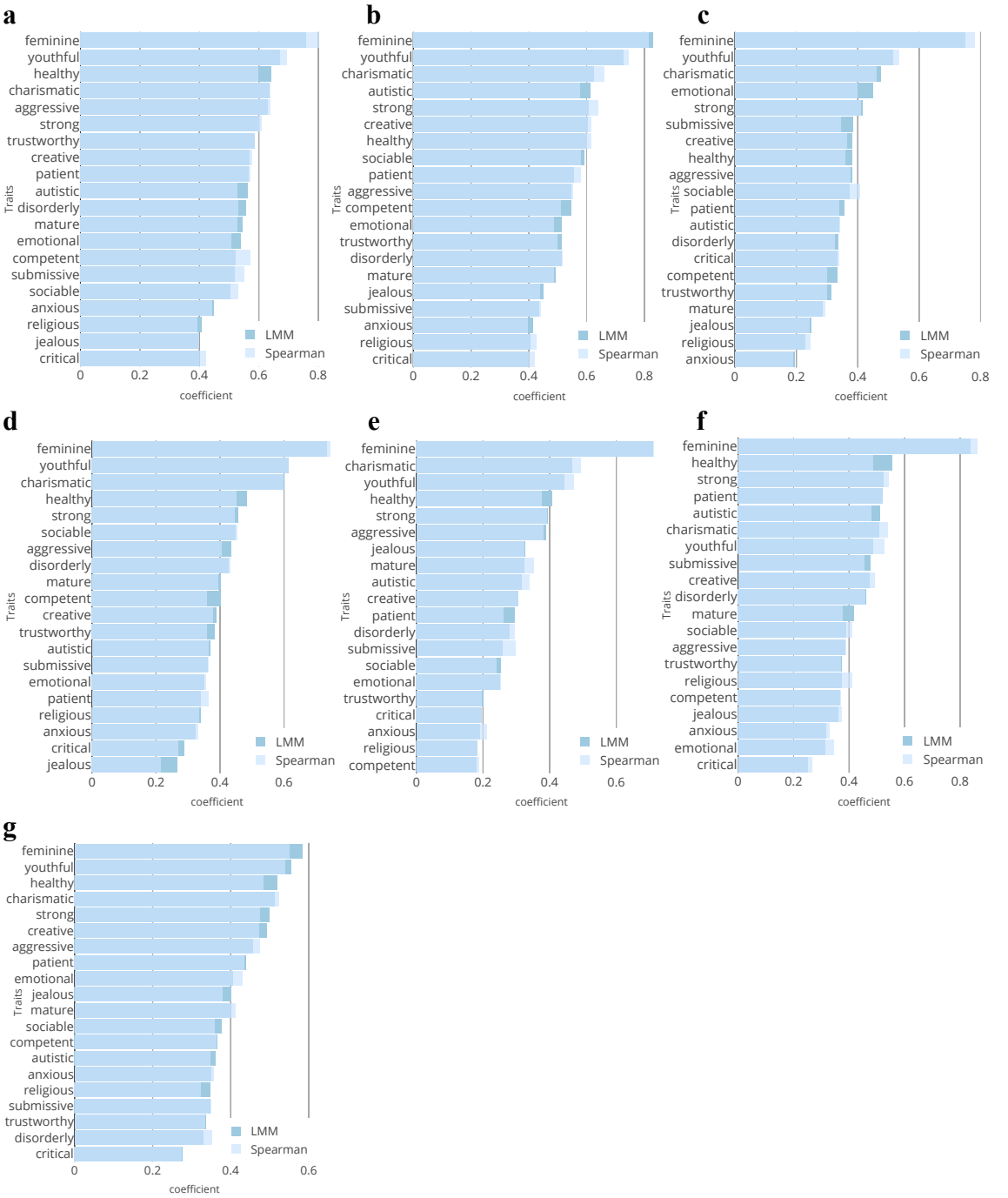


Supplementary Figure 4: Correlations among trait attributions from faces. Panel (a) plots the Pearson correlation matrix of the one hundred trait attributions. Panel (b) plots the average (red), median (blue), and maximum (purple) correlations a trait has with all the other ninety-nine traits. The horizontal dashed line indicates $r = 0.30$. To the far left are the 8 traits we excluded from EFA because of their low average correlations with all other traits (points below the dashed line: *sarcastic*, *white*, *thrifty*, *shallow*, *homosexual*, *nosey*, *conservative*, and *reserved*).

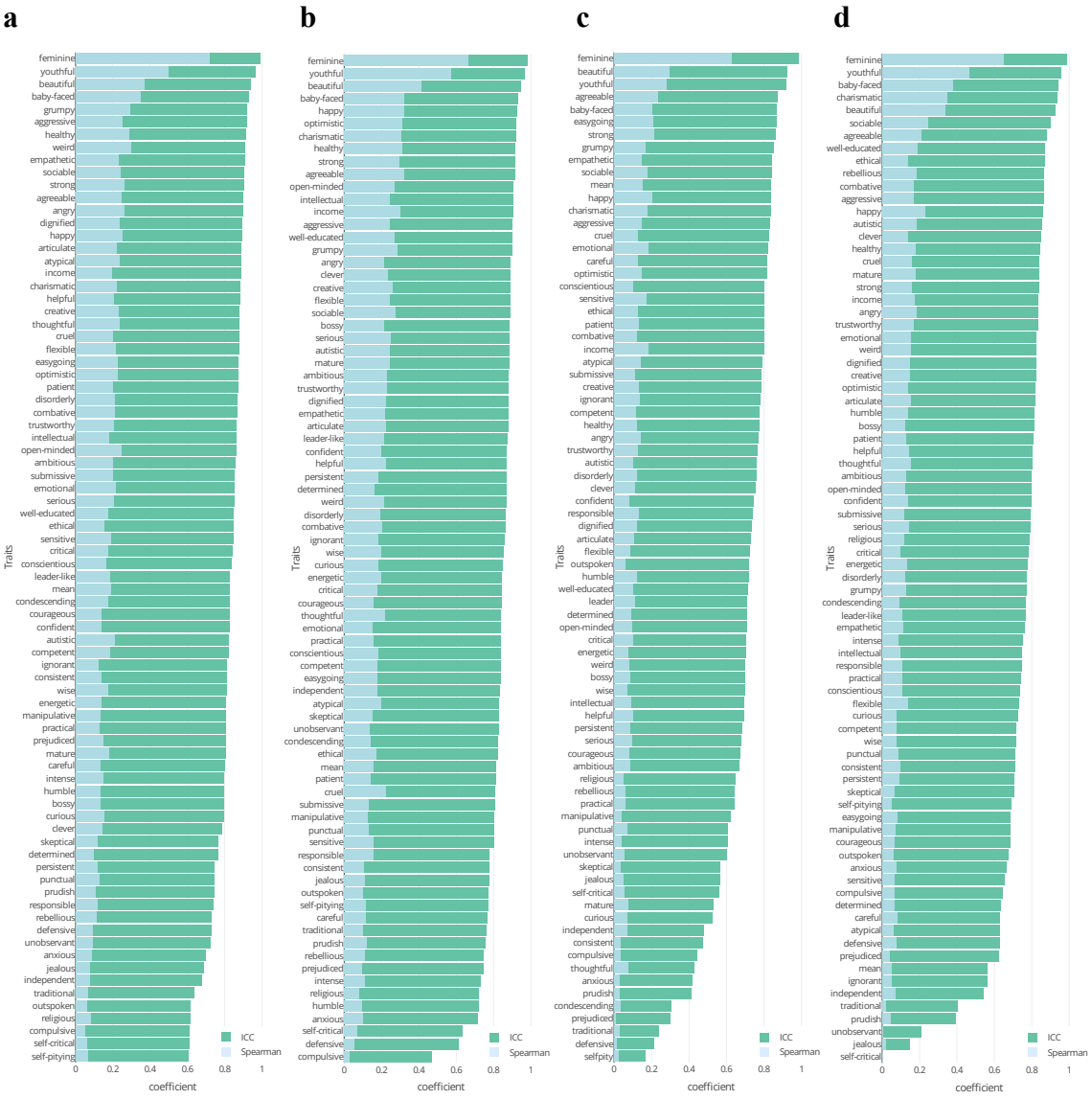


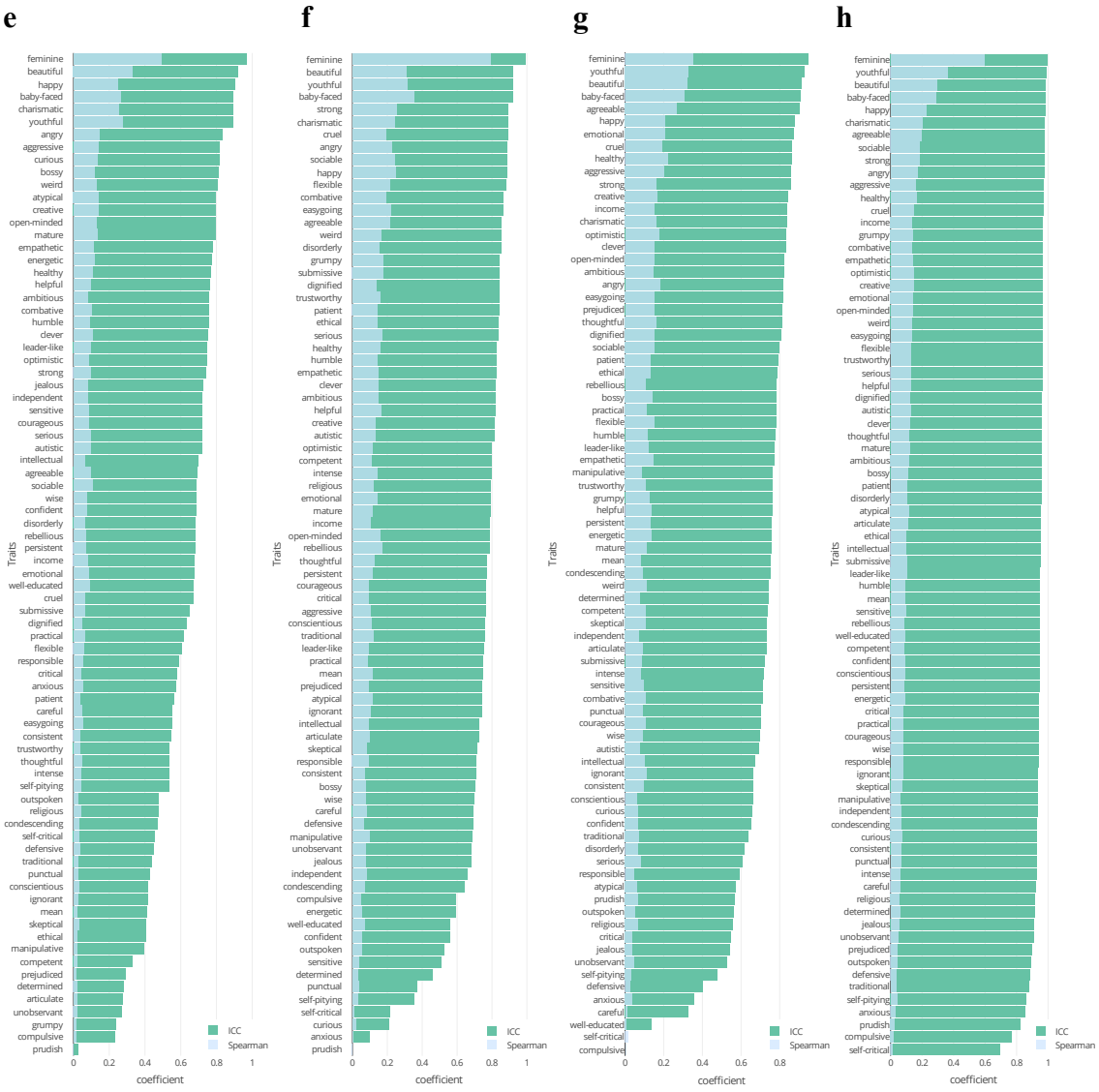
Supplementary Figure 5: Eigenvalue plots and results of five methods for determining the optimal number of factors. Panel (a) plots the results for Study 1 data which consist of aggregate ratings for the ninety-two traits used for the final EFA. Panel (b) plots the

results for Study 1 data which consist of aggregate ratings for all one hundred traits (i.e., without excluding the eight low-factorability traits). The other panels plot the results for Study 2 data from North America (c), Latvia (d), Peru (e), the Philippines (f), India (g), Kenya (h), and Gaza (i). As recommended by previous research, we applied parallel analysis to determine the optimal number of factors to retain. Parallel analysis retains factors that are not simply due to chance by comparing the eigenvalues of the observed data matrix with those of multiple randomly generated data matrices that match the sample size of the observed data matrix. For comparison, we also obtained estimations based on Kaiser's rule, which retains factors with eigenvalues that are greater than one, Cattell's scree test, which retains factors to the left of the point from which the plotted ordered eigenvalues could be approximated with a straight line, the optimal coordinates index, which provides a non-graphical solution to Cattell's scree test based on the linear extrapolation, and the acceleration factor, which provides a non-graphical solution to Cattell's scree test based on the second derivative.



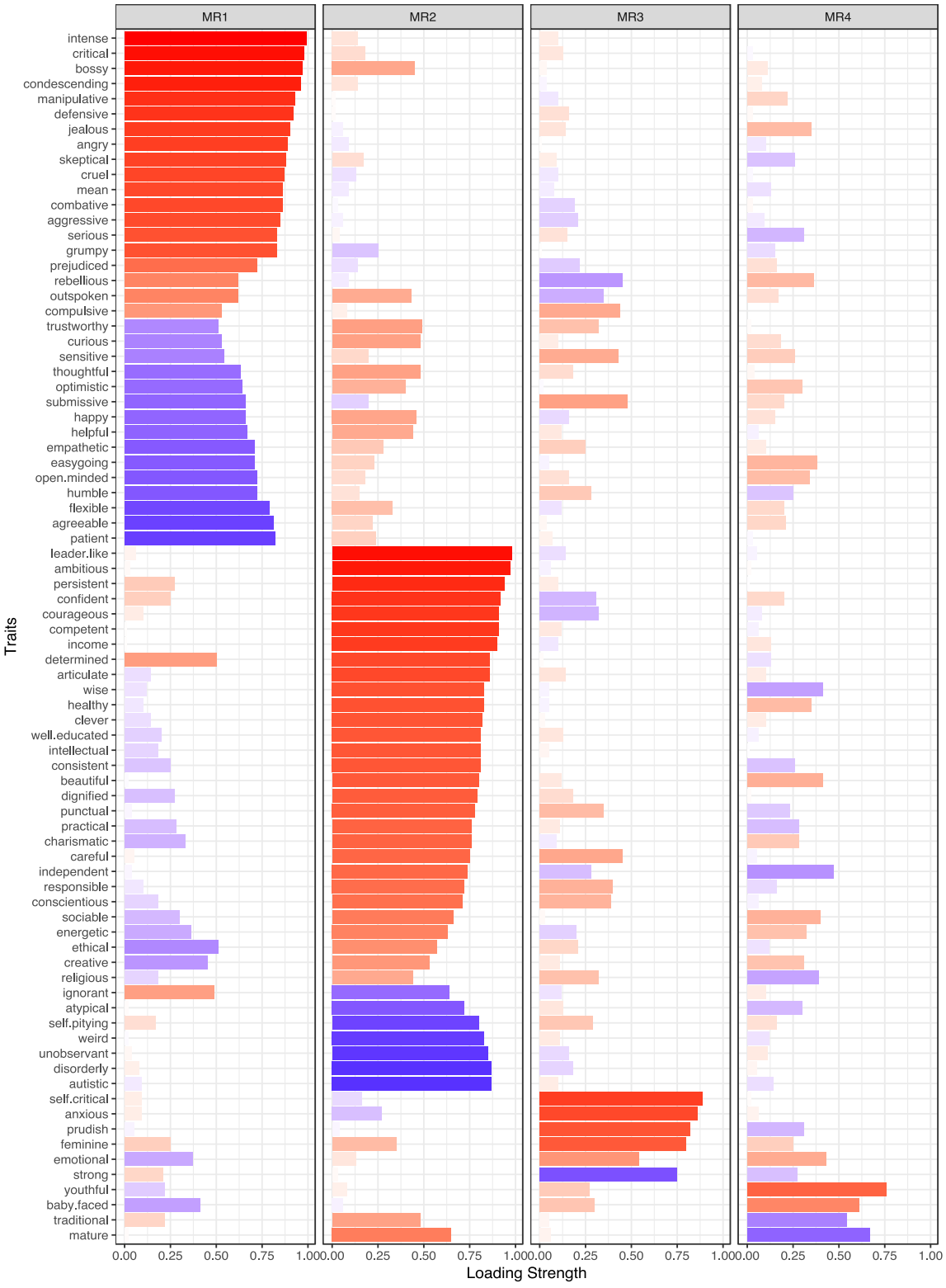
Supplementary Figure 6: Within-subject test-retest reliability of trait attributions from faces in seven countries. All participants in all countries rated a subset of twenty traits twice for all the one hundred faces. Each panel plots results from one sample in North America (a), Latvia (b), Peru (c), the Philippines (d), India (e), Kenya (f), and Gaza (g), respectively. Within-subject test-retest reliability was assessed with two methods, linear mixed-effect modeling (dark blue) and Spearman correlations (light blue). Each row indicates one trait. The length of the bar indicates the magnitude of the coefficient.


a
b
c
d
Traits
coefficient
ICC
Spearman
0
0.2
0.4
0.6
0.8
1

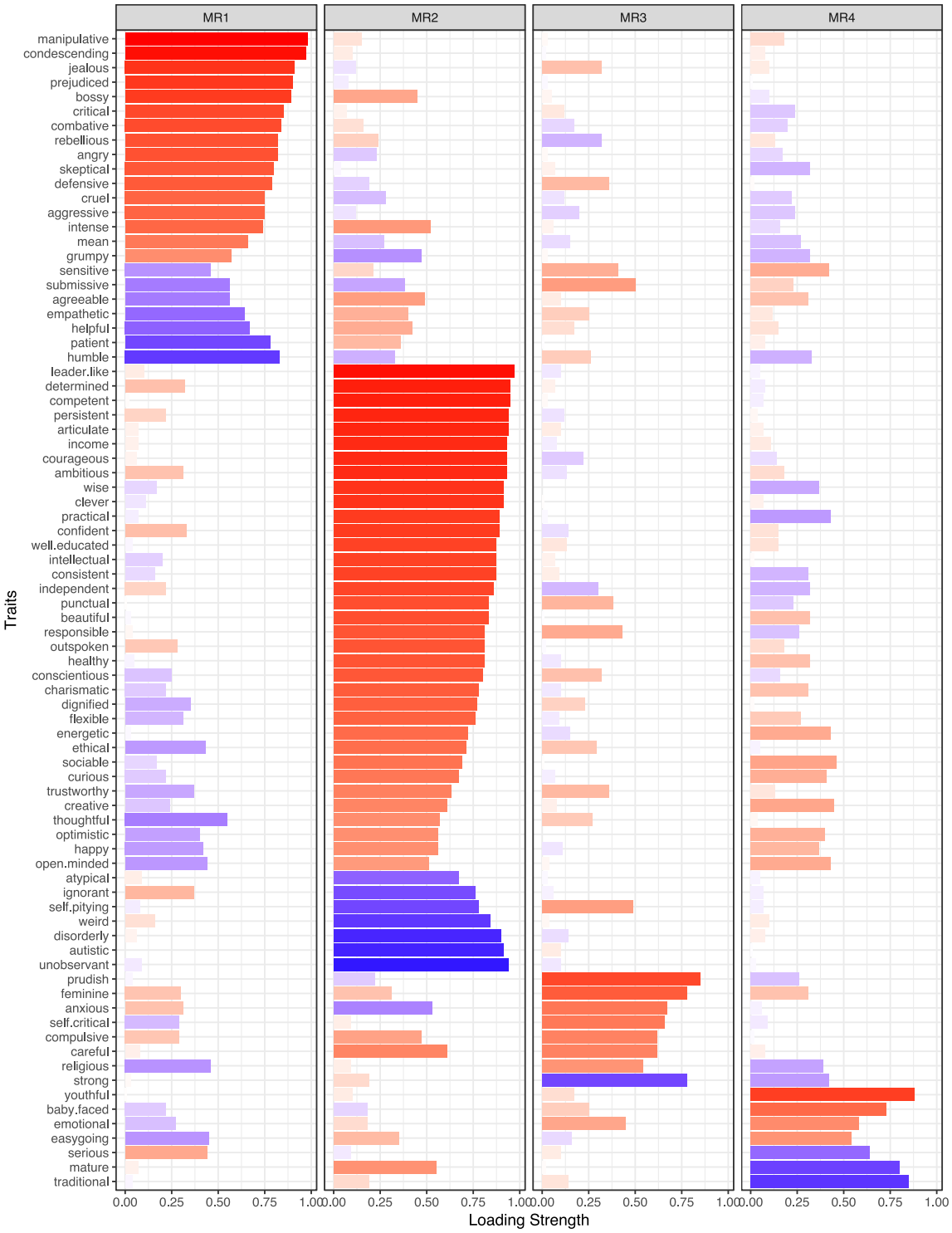


Supplementary Figure 7: Between-subject consensus of trait attributions from faces in each country and across all countries. The first seven panels each plots results for North America (a), Latvia (b), Peru (c), the Philippines (d), India (e), Kenya (f), and Gaza (g), respectively. Panel (h) plots results across the seven locations. Between-subject consensus was assessed with two methods, intraclass correlation coefficients (dark green) and Spearman correlations (light green). Each row indicates results for one trait. The length of a bar indicates the magnitude of the coefficient.

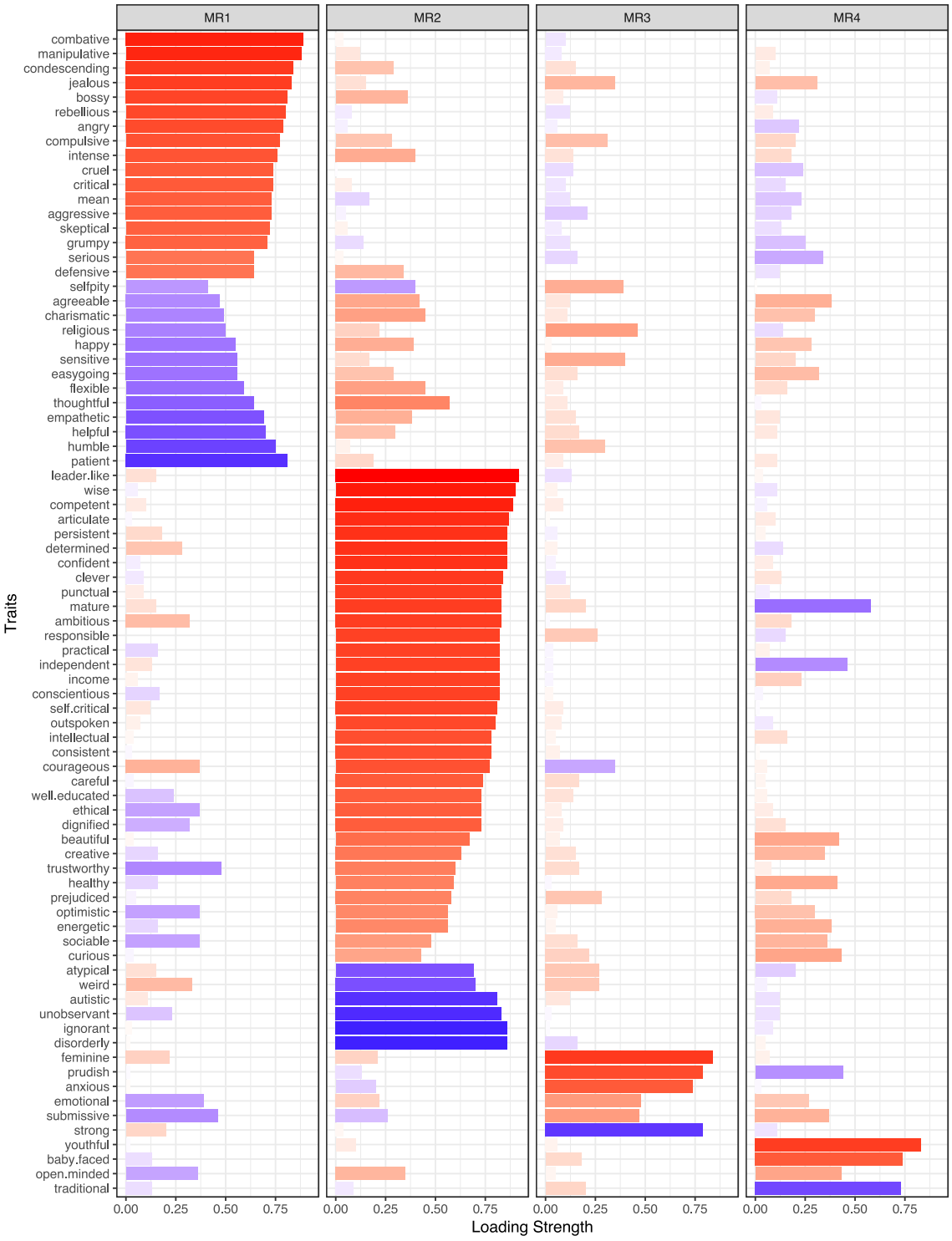
a


MR1
MR2
MR3
MR4
intense
critical
bossy
condescending
manipulative
defensive
jealous
angry
skeptical
cruel
mean
combative
aggressive
serious
grumpy
prejudiced
rebellious
outspoken
compulsive
trustworthy
curious
sensitive
thoughtful
optimistic
submissive
happy
helpful
empathetic
easygoing
open.minded
humble
flexible
agreeable
patient
leader.like
ambitious
persistent
confident
courageous
competent
income
determined
articulate
wise
healthy
clever
well.educated
intellectual
consistent
beautiful
dignified
punctual
practical
charismatic
careful
independent
responsible
conscientious
sociable
energetic
ethical
creative
religious
ignorant
atypical
self.pitying
weird
unobservant
disorderly
autistic
self.critical
anxious
prudish
feminine
emotional
strong
youthful
baby.faced
traditional
mature
Traits
0.00
0.25
0.50
0.75
1.00
Loading Strength

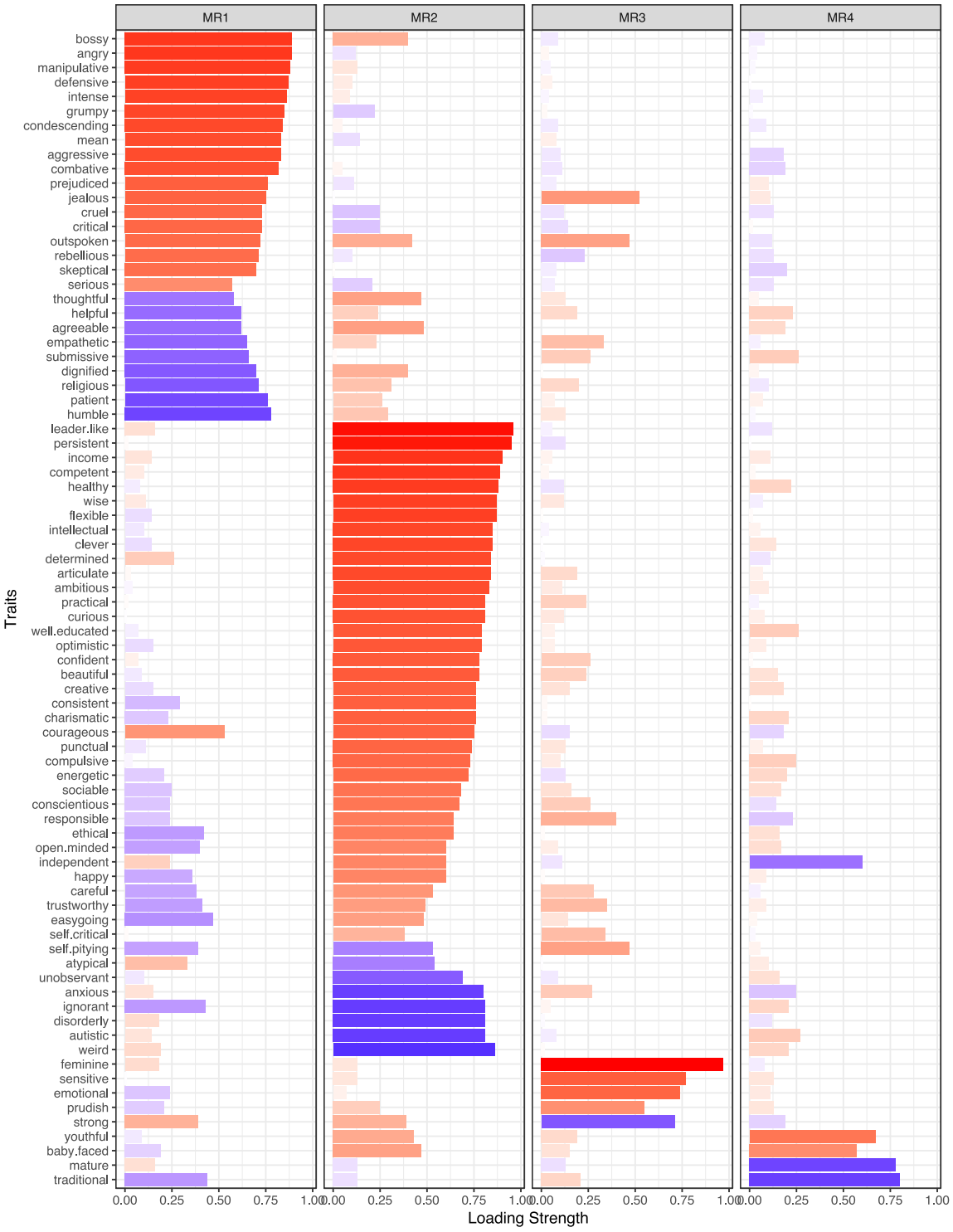
b


MR1
MR2
MR3
MR4
Traits
manipulative
condescending
jealous
prejudiced
bossy
critical
combative
rebellious
angry
skeptical
defensive
cruel
aggressive
intense
mean
grumpy
sensitive
submissive
agreeable
empathetic
helpful
patient
humble
leader.like
determined
competent
persistent
articulate
income
courageous
ambitious
wise
clever
practical
confident
well.educated
intellectual
consistent
independent
punctual
beautiful
responsible
outspoken
healthy
conscientious
charismatic
dignified
flexible
energetic
ethical
sociable
curious
trustworthy
creative
thoughtful
optimistic
happy
open.minded
atypical
ignorant
self.pitying
weird
disorderly
autistic
unobservant
prudish
feminine
anxious
self.critical
compulsive
careful
religious
strong
youthful
baby.faced
emotional
easygoing
serious
mature
traditional
0.00
0.25
0.50
0.75
1.00
Loading Strength

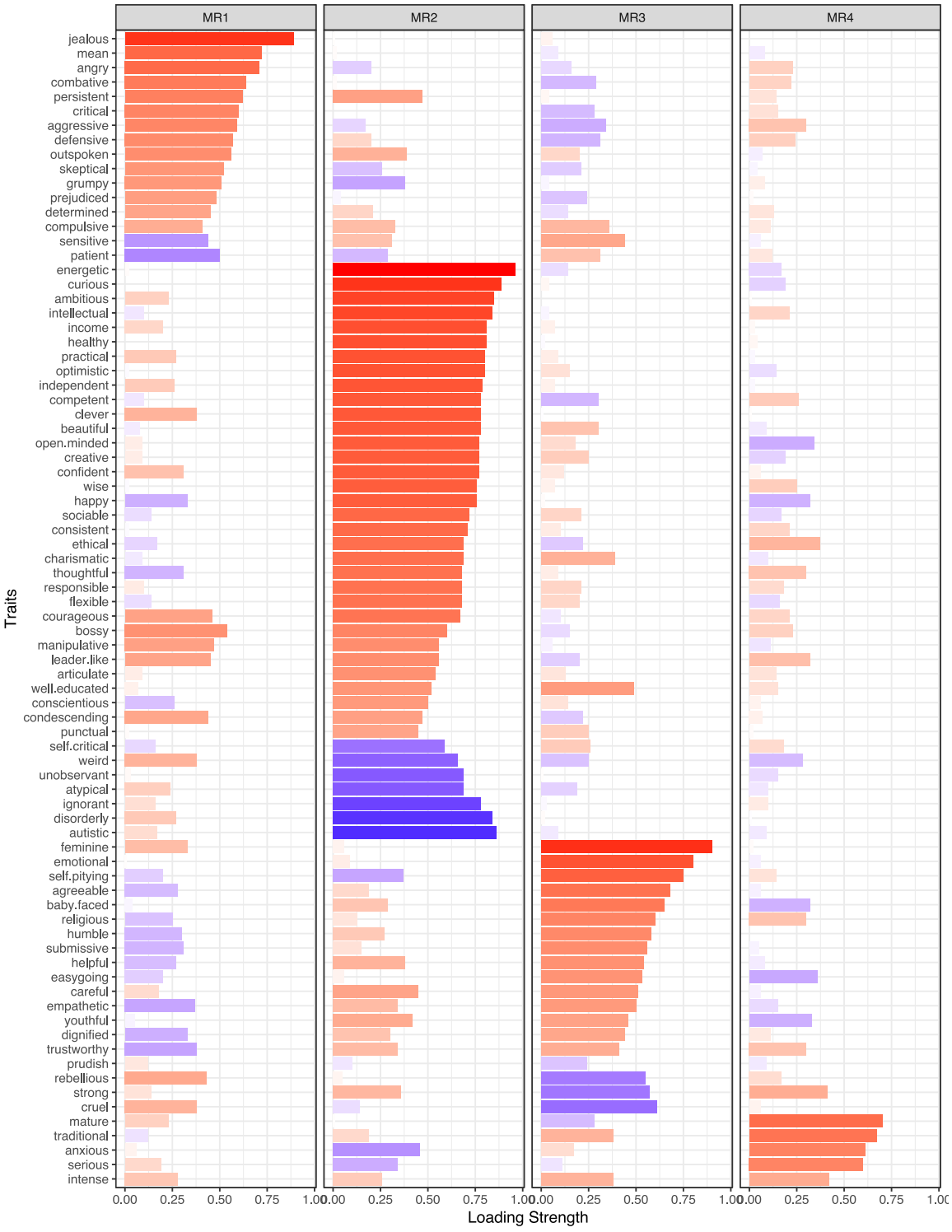
c


MR1
MR2
MR3
MR4
Traits
combative
manipulative
condescending
jealous
bossy
rebellious
angry
compulsive
intense
cruel
critical
mean
aggressive
skeptical
grumpy
serious
defensive
selfpity
agreeable
charismatic
religious
happy
sensitive
easygoing
flexible
thoughtful
empathetic
helpful
humble
patient
leader.like
wise
competent
articulate
persistent
determined
confident
clever
punctual
mature
ambitious
responsible
practical
independent
income
conscientious
self.critical
outspoken
intellectual
consistent
courageous
careful
well.educated
ethical
dignified
beautiful
creative
trustworthy
healthy
prejudiced
optimistic
energetic
sociable
curious
atypical
weird
autistic
unobservant
ignorant
disorderly
feminine
prudish
anxious
emotional
submissive
strong
youthful
baby.faced
open.minded
traditional
0.00
0.25
0.50
0.75
Loading Strength

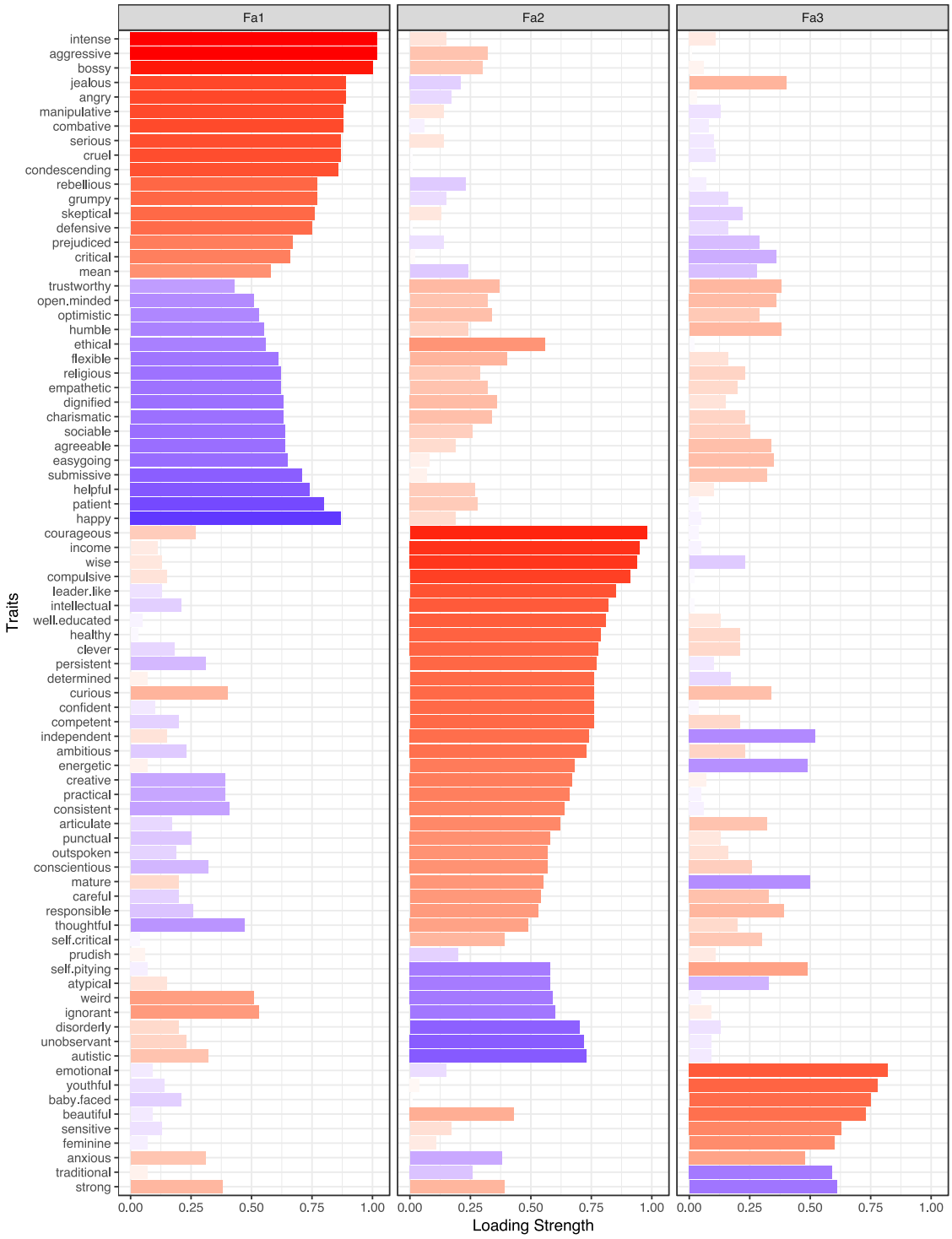
d


MR1
MR2
MR3
MR4
Traits
bossy
angry
manipulative
defensive
intense
grumpy
condescending
mean
aggressive
combative
prejudiced
jealous
cruel
critical
outspoken
rebellious
skeptical
serious
thoughtful
helpful
agreeable
empathetic
submissive
dignified
religious
patient
humble
leader.like
persistent
income
competent
healthy
wise
flexible
intellectual
clever
determined
articulate
ambitious
practical
curious
well.educated
optimistic
confident
beautiful
creative
consistent
charismatic
courageous
punctual
compulsive
energetic
sociable
conscientious
responsible
ethical
open.minded
independent
happy
careful
trustworthy
easygoing
self.critical
self.pitying
atypical
unobservant
anxious
ignorant
disorderly
autistic
weird
feminine
sensitive
emotional
prudish
strong
youthful
baby.faced
mature
traditional
0.00
0.25
0.50
0.75
1.00
Loading Strength

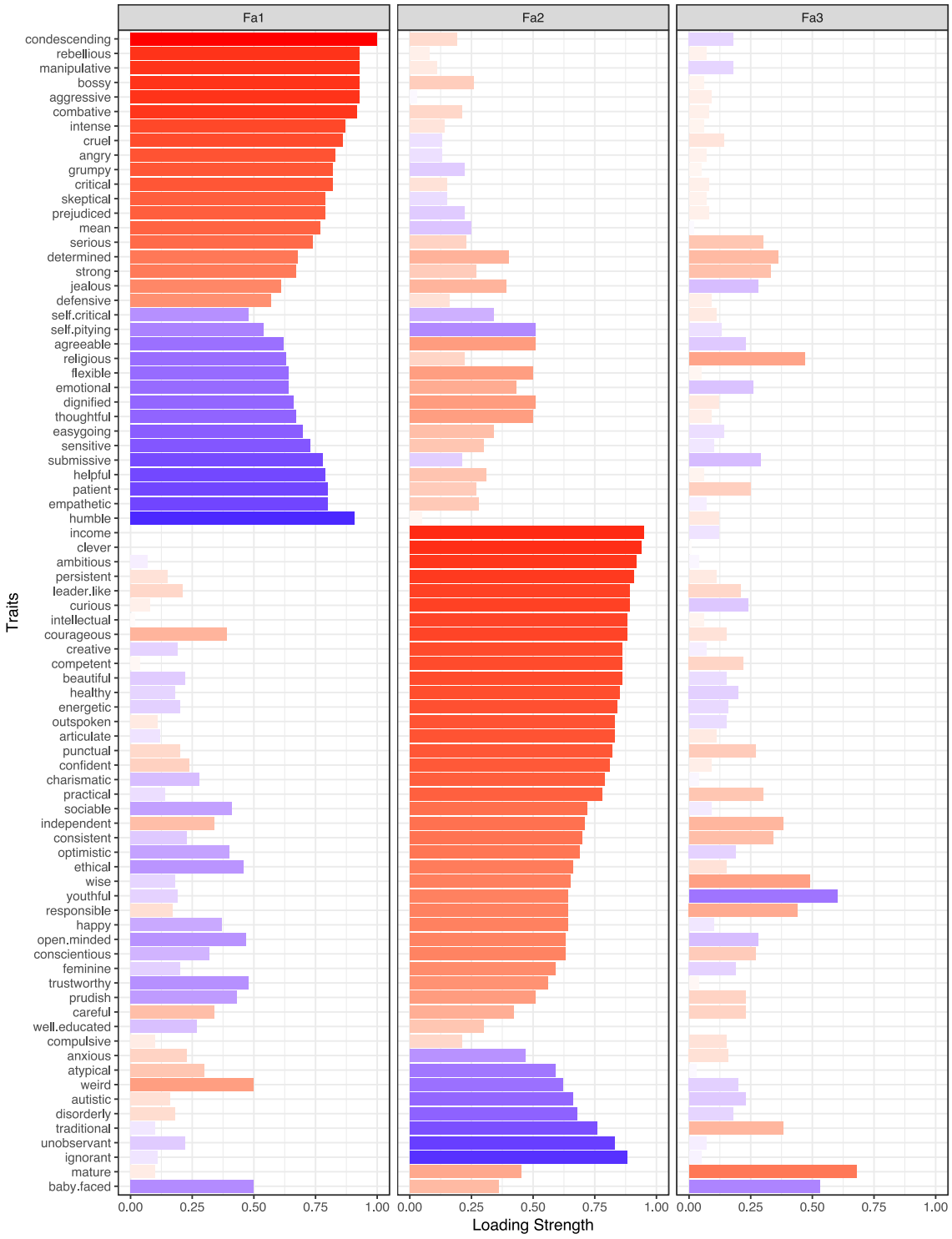
e


MR1
MR2
MR3
MR4
jealous
mean
angry
combative
persistent
critical
aggressive
defensive
outspoken
skeptical
grumpy
prejudiced
determined
compulsive
sensitive
patient
energetic
curious
ambitious
intellectual
income
healthy
practical
optimistic
independent
competent
clever
beautiful
open.minded
creative
confident
wise
happy
sociable
consistent
ethical
charismatic
thoughtful
responsible
flexible
courageous
bossy
manipulative
leader.like
articulate
well.educated
conscientious
condescending
punctual
self.critical
weird
unobservant
atypical
ignorant
disorderly
autistic
feminine
emotional
self.pitying
agreeable
baby.faced
religious
humble
submissive
helpful
easygoing
careful
empathetic
youthful
dignified
trustworthy
prudish
rebellious
strong
cruel
mature
traditional
anxious
serious
intense
Traits
0.00
0.25
0.50
0.75
1.00
Loading Strength

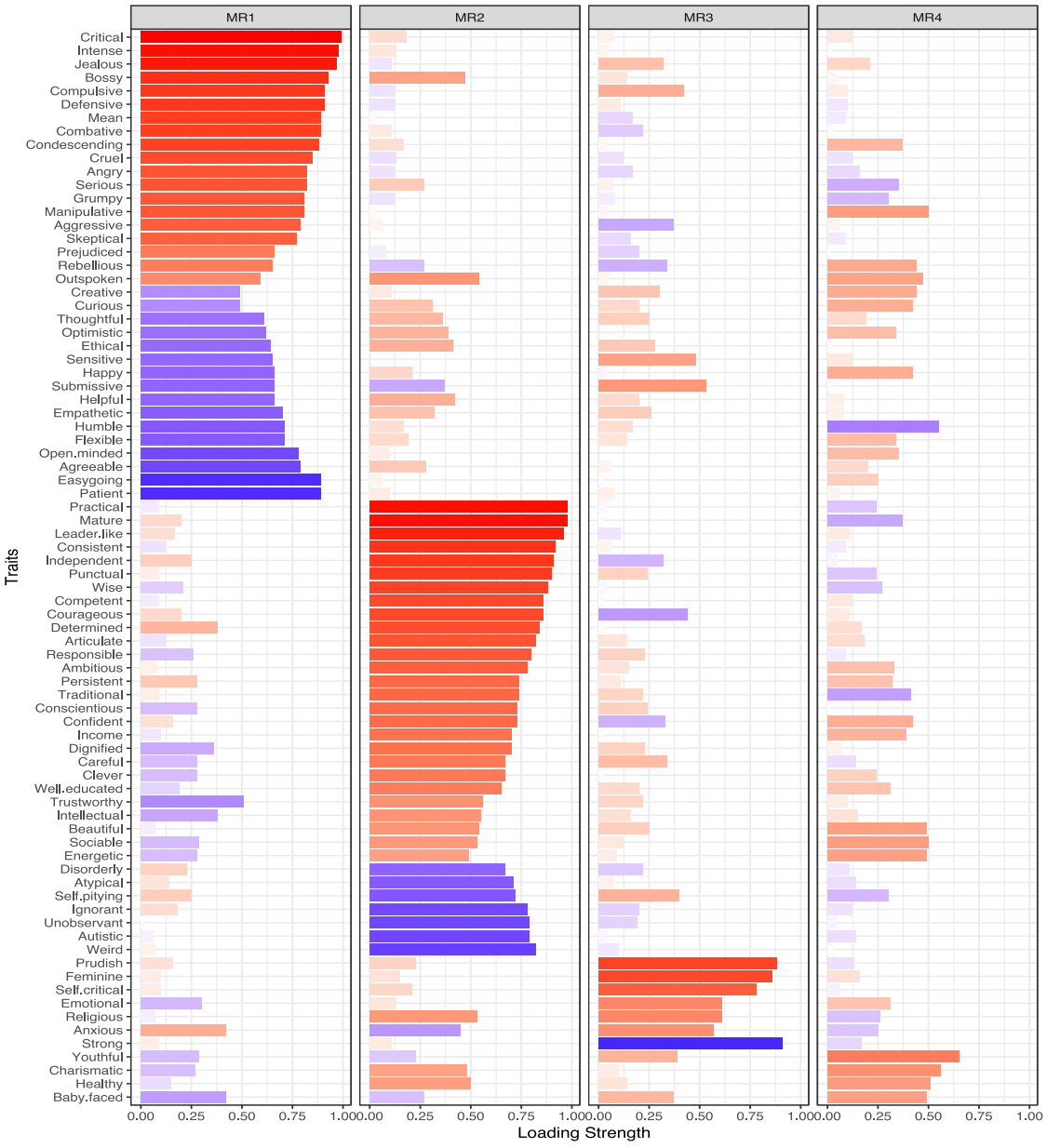
f


Fa1
Fa2
Fa3
Traits
intense
aggressive
bossy
jealous
angry
manipulative
combative
serious
cruel
condescending
rebellious
grumpy
skeptical
defensive
prejudiced
critical
mean
trustworthy
open.minded
optimistic
humble
ethical
flexible
religious
empathetic
dignified
charismatic
sociable
agreeable
easygoing
submissive
helpful
patient
happy
courageous
income
wise
compulsive
leader.like
intellectual
well.educated
healthy
clever
persistent
determined
curious
confident
competent
independent
ambitious
energetic
creative
practical
consistent
articulate
punctual
outspoken
conscientious
mature
careful
responsible
thoughtful
self.critical
prudish
self.pitying
atypical
weird
ignorant
disorderly
unobservant
autistic
emotional
youthful
baby.faced
beautiful
sensitive
feminine
anxious
traditional
strong
0.00
0.25
0.50
0.75
1.00
Loading Strength

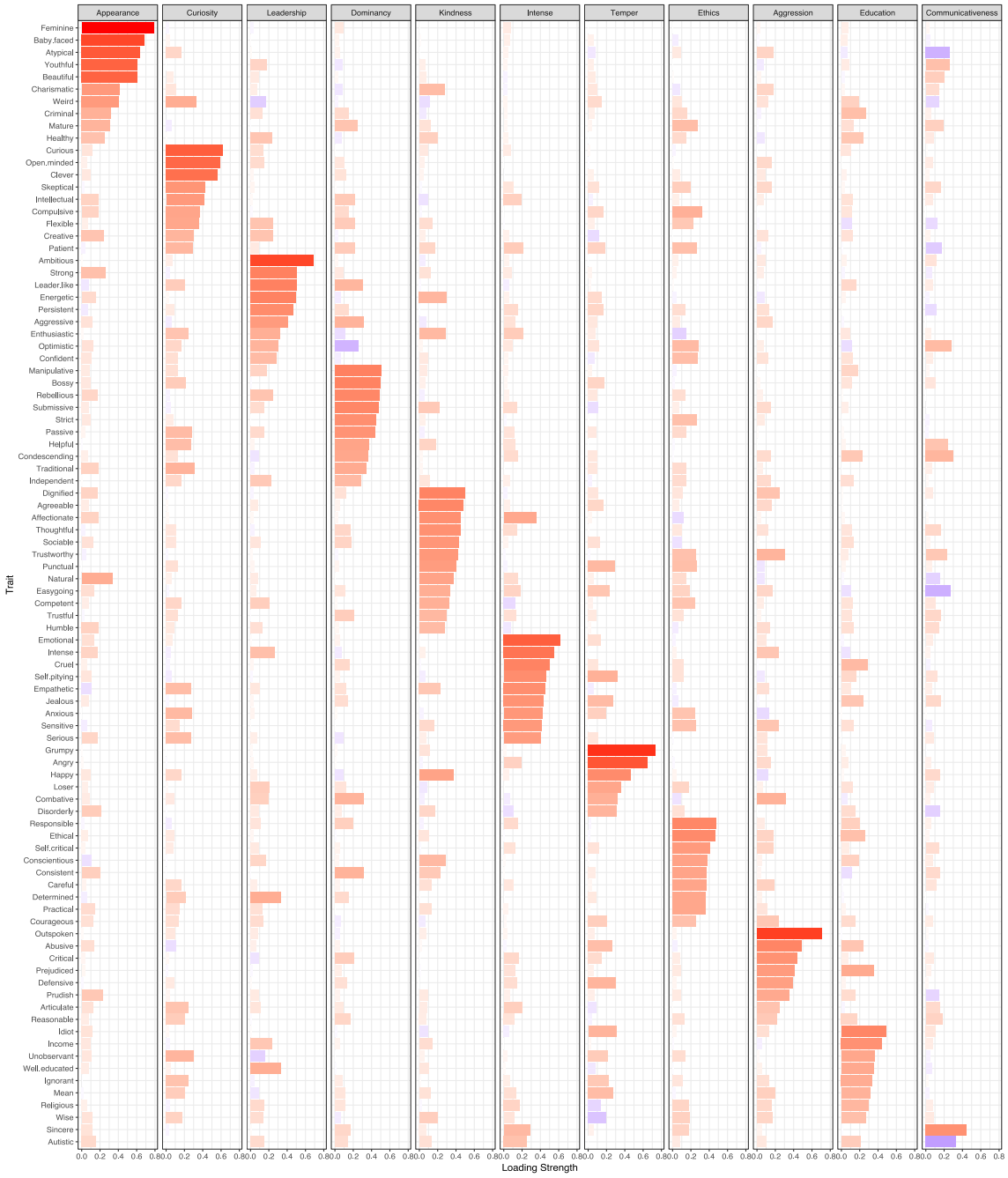
g


Fa1
Fa2
Fa3
condescending
rebellious
manipulative
bossy
aggressive
combative
intense
cruel
angry
grumpy
critical
skeptical
prejudiced
mean
serious
determined
strong
jealous
defensive
self.critical
self.pitying
agreeable
religious
flexible
emotional
dignified
thoughtful
easygoing
sensitive
submissive
helpful
patient
empathetic
humble
income
clever
ambitious
persistent
leader.like
curious
intellectual
courageous
creative
competent
beautiful
healthy
energetic
outspoken
articulate
punctual
confident
charismatic
practical
sociable
independent
consistent
optimistic
ethical
wise
youthful
responsible
happy
open.minded
conscientious
feminine
trustworthy
prudish
careful
well.educated
compulsive
anxious
atypical
weird
autistic
disorderly
traditional
unobservant
ignorant
mature
baby.faced
Traits
0.00
0.25
0.50
0.75
1.00
Loading Strength

h

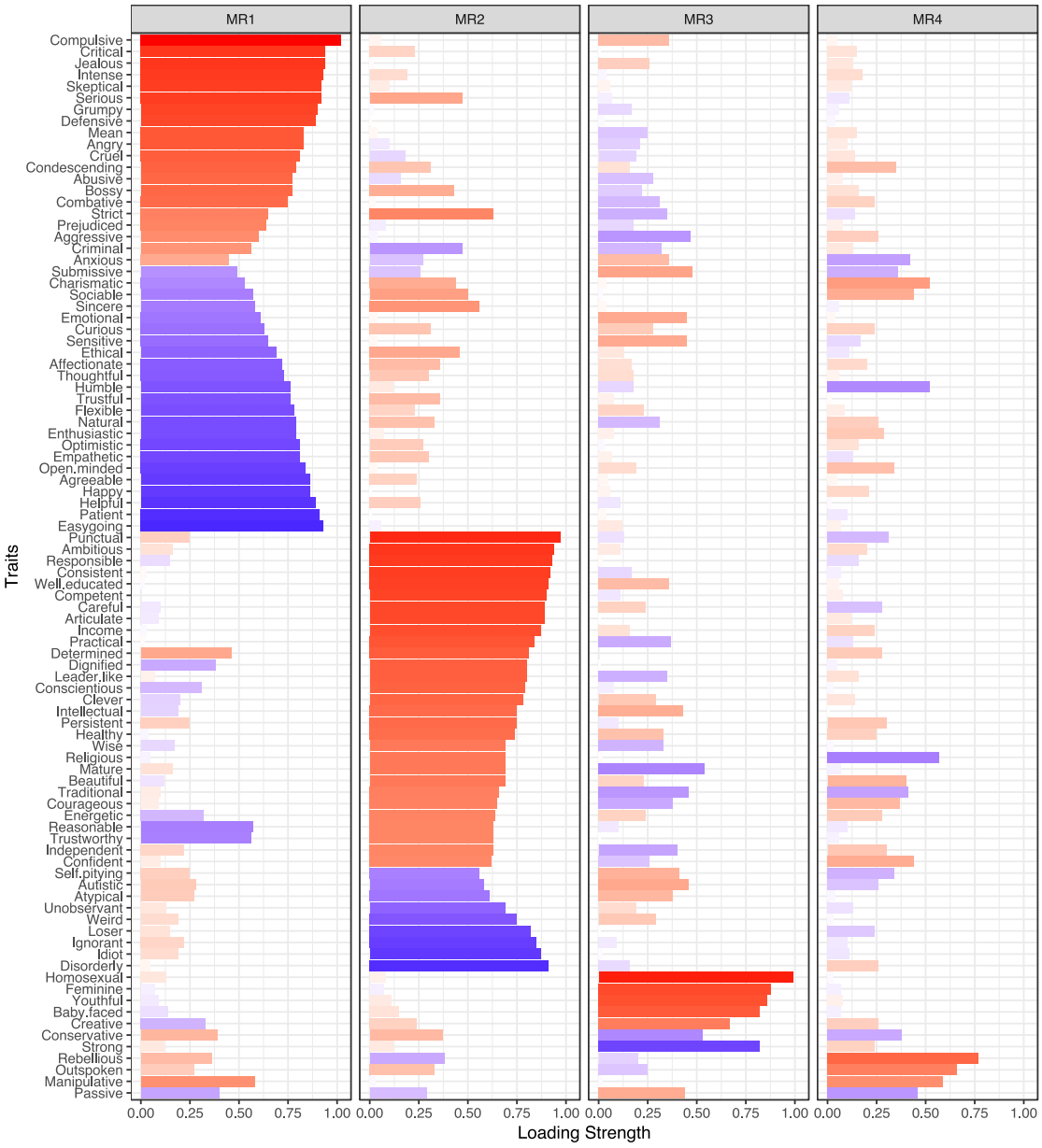


Supplementary Figure 8: Standardized factor loadings of 80 traits. The first seven panels plot results for North America (a), Latvia (b), Peru (c), the Philippines (d), India (e), Kenya (f), and Gaza (g) in Study 2, respectively. Factors within a panel for each sample were reordered (if needed) to highlight their correspondence with the four dimensions found in Study 1. The last panel (h) plots the results for the subset of data from Study 1 that consists of aggregate ratings for the same 80 traits as in Study 2. Each column plots the strength of the factor loadings across the 80 traits. The color of the bar indicates the sign of the loading (red: positive; blue: negative); the length and saturation of the bar indicate the magnitude of the loading.

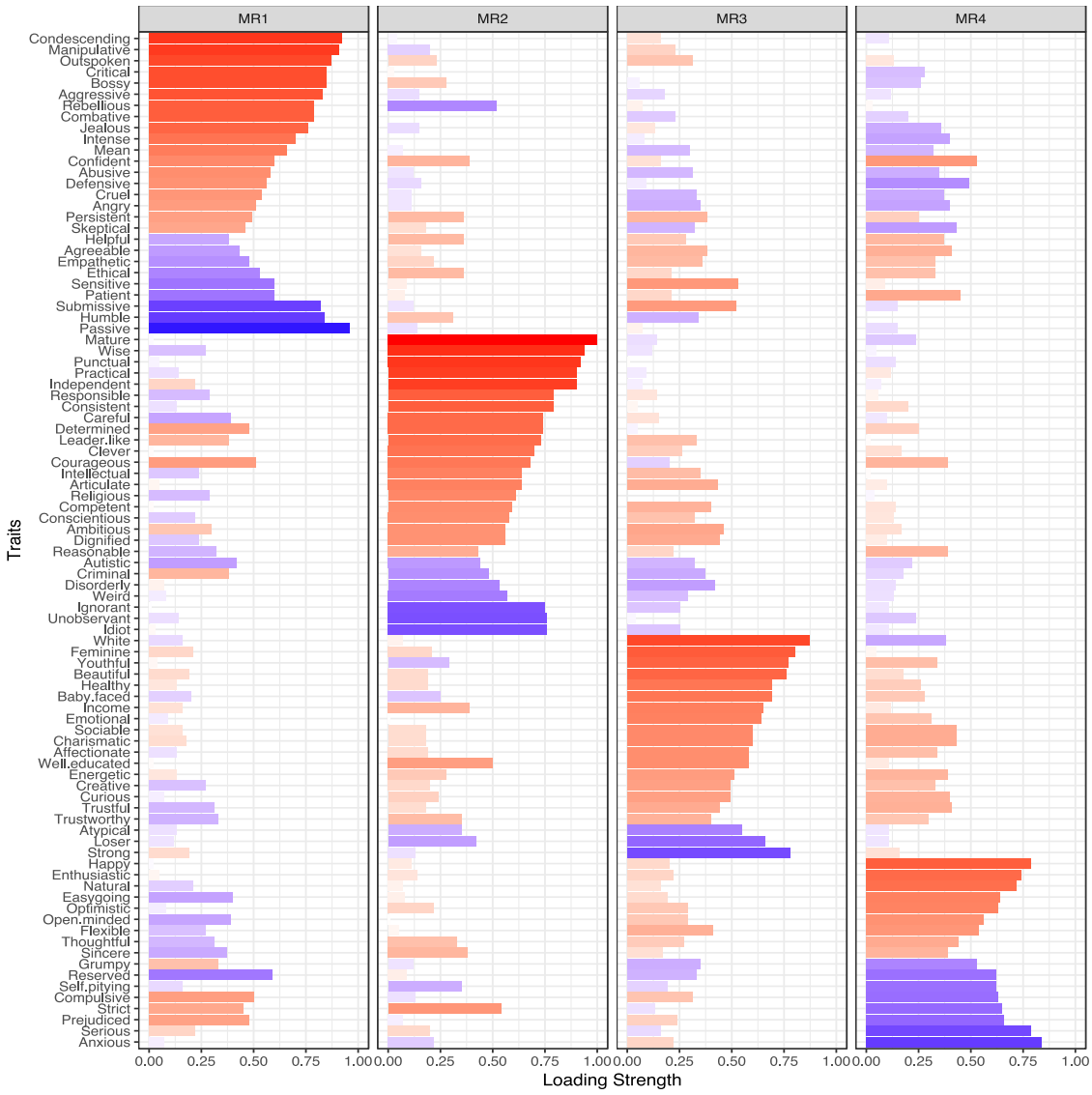


Supplementary Figure 9: Standardized factor loadings of the 92 traits from exploratory factor analysis based on their semantic similarities. Each column plots the strength of the factor loadings across the 92 traits. The color of the bar indicates the sign of the loading (red: positive; blue: negative). The length and saturation of the bar indicate the strength of the loading (the absolute value of the loading). The eleven semantic factors each accounted for 8%, 8%, 7%, 7%, 6%, 7%, 6%, 7%, 6%, 6%, 3% of the common variance and were moderately correlated (correlations ranged from 0.16 to 0.43).

a


MR1
MR2
MR3
MR4
Traits
Compulsive
Critical
Jealous
Intense
Skeptical
Serious
Grumpy
Defensive
Mean
Angry
Cruel
Condescending
Abusive
Bossy
Combative
Strict
Prejudiced
Aggressive
Criminal
Anxious
Submissive
Charismatic
Sociable
Sincere
Emotional
Curious
Sensitive
Ethical
Affectionate
Thoughtful
Humble
Trustful
Flexible
Natural
Enthusiastic
Optimistic
Empathetic
Open.minded
Agreeable
Happy
Helpful
Patient
Easygoing
Punctual
Ambitious
Responsible
Consistent
Well.educated
Competent
Careful
Articulate
Income
Practical
Determined
Dignified
Leader.like
Conscientious
Clever
Intellectual
Persistent
Healthy
Wise
Religious
Mature
Beautiful
Traditional
Courageous
Energetic
Reasonable
Trustworthy
Independent
Confident
Self.pitying
Autistic
Atypical
Unobservant
Weird
Loser
Ignorant
Idiot
Disorderly
Homosexual
Feminine
Youthful
Baby.faced
Creative
Conservative
Strong
Rebellious
Outspoken
Manipulative
Passive
0.00
0.25
0.50
0.75
1.00
Loading Strength

b



Supplementary Figure 10: Standardized factor loadings based on data for male faces (a) and female faces (b). Each column plots the strength of the factor loadings across all 92 traits. The color of the bar indicates the sign of the loading (red: positive; blue: negative). The length and saturation of the bar indicate the strength of the loading (the absolute value of the loading). Eight of the one hundred traits were excluded from dimensional analyses for the male faces due to their low factorability. The four factors for the male faces each accounted for 34%, 31%, 13%, and 7% of the common variance and were weakly correlated (r_{12} = -0.24, r_{13} = -0.37, r_{14} = 0.10, r_{23} = -0.14, r_{24} = 0.22, r_{34} = -0.10). Nine of the one hundred traits were excluded from dimensional analyses for the female faces due to their low factorability. The four factors for the female faces each accounted for 21%, 23%, 21%, and 20% of the common variance and were moderately correlated (r_{12} = -0.05, r_{13} = -0.15, r_{14} = -0.30, r_{23} = 0.42, r_{24} = 0.39, r_{34} = 0.55).

Data-driven Four Dimensions	Five Factor1	Five Factor2	Five Factor3	Five Factor4	Five Factor5	Three Factor1	Three Factor2	Three Factor3
Critical/condescending	**0.99**	-0.2	-0.26	-0.04	-0.33	**-0.98**	-0.11	-0.21
Leadership/competence	-0.14	**0.95**	0.29	0.55	-0.28	0.15	**0.96**	0.28
Female-stereotype	-0.23	0.22	**0.97**	-0.08	0.26	0.34	0.19	**0.93**
Youth-stereotype	-0.22	0.40	0.21	**0.76**	**0.70**	0.42	0.54	-0.13

Supplementary Table 1: Tucker indices of factor congruence between different numbers of factor solutions. The rows list our four data-driven dimensions of trait attributions from faces. Factors in the first five columns were obtained when a five-factor solution was forced to explain the common variance in the data. Factors in the last three columns were obtained when a three-factor solution was forced to explain the common variance in the data. A factor's congruence indices with the four data-driven dimensions are listed in each column with the highest value highlighted in bold.

	Factor 1	Factor 2	Factor 3
Wise	**0.92**	-0.37	0.02
Trustworthy	**0.80**	0.2	0.24
Agreeable	**0.68**	0.2	0.43
Confident	**0.63**	0.13	-0.63
Happy	**0.61**	0.21	0.26
Beautiful	**0.60**	0.54	-0.23
Feminine	**0.31**	0.28	0.2
Youthful	-0.11	**0.98**	0.12
Baby-faced	-0.09	**0.82**	0.31
Healthy	0.52	**0.67**	-0.25
White	0.16	**0.27**	0.05
Submissive	0.05	0.21	**0.88**
Aggressive	-0.38	-0.12	**-0.79**

Supplementary Table 2: Standardized factor loadings of the 13 traits. Each column lists the factor loadings across the 13 traits that matched those used in Sutherland et al. (2013) and that were included in the trait inventory of the present research. For each trait, its highest loading across the three factors is highlighted in bold.

Supplementary Methods

Traits	Definitions
PERSONALITY TRAITS	
affectionate	A person who is comfortable showing his/her love, warmth, and kindness
sincere	A person who says what he/she genuinely feels or believes
helpful	A person who gives help when others are in need
sensitive	A person who is aware of or careful about others' attitudes, feelings, or circumstances
agreeable	A person who is kind, cooperative, and sympathetic
feminine	A person whose facial appearance looks like a woman
trustful	A person who tends to trust other people easily (note: this is different from being trustworthy)
thoughtful	A person who is considerate of others' needs
flexible	A person who is ready and able to change so as to adapt to different circumstances
reasonable	A person who makes sense and whose opinions most people would agree with
humble	A person who is modest and does not boast
religious	A person who practices religion and believes in their faith
optimistic	A person who is hopeful and confident about the future
conscientious	A person who does his/her work or duty thoroughly and responsibly
natural	A person who is relaxed and spontaneous
abusive	A person who is extremely offensive and insulting
combative	A person who likes to argue or pick a fight
cruel	A person who willfully causes pain or suffering to other people or to animals, and feels no concern about it
critical	A person who judges others harshly, and often makes disapproving comments
rebellious	A person who resists authority, control, or convention and wants to have their own way
sarcastic	A person who likes using irony in order to mock others
prejudiced	A person who holds biased judgments about other people; bigoted
manipulative	A person who likes to control people in order to meet his/her own needs
skeptical	A person who questions things and is not easily convinced
aggressive	A person who pursues his/her aims and interests forcefully, sometimes with physical force
sociable	A person who is friendly and enjoys talking and engaging in activities with other people
enthusiastic	A person who is filled with eager enjoyment and interest
confident	A person who is sure about his/her own abilities, correctness, and successfulness

energetic	A person who is very active and full of energy
outspoken	A person who is frank in stating his/her opinions especially if they are critical or controversial
persistent	A person who is able to continue in a course of action in spite of difficulty or opposition
reserved	A person who tends not to show their emotions or opinions and is quiet
passive	A person who allows things to happen or accepts what others do, without resistance or trying to change anything
submissive	A person who shows a willingness to be controlled by others or conforms to the authority or will of others
serious	A person who shows deep thoughts and who doesn't smile or laugh easily
prudish	A person who is overly proper and cannot stand hearing any sexual reference
responsible	A person who accepts the consequences of his or her own actions and decisions
practical	A person who is sensible and realistic in dealing with a situation or problem
careful	A person who works and thinks in a cautious, thorough, or thoughtful way to avoid potential danger
consistent	A person who behaves or responds in the same way over time; reliable
punctual	A person who is always on time
strict	A person who follows rules exactly, and expects others to follow rules exactly
dignified	A person who is polite and composed, and always shows good and respected manners
mature	A person who thinks and behaves like a responsible adult
thrifty	A person who uses money and other resources carefully and not wastefully
ambitious	A person who has a strong desire and determination to succeed in their goals
conservative	A person who sticks to traditional values, especially in politics or religion, and who does not like new ideas or changes
wise	A person who has mature experience, knowledge, and good judgments
disorderly	A person who is untidy and not organized
unobservant	A person who does not notice things
emotional	A person who shows his/her feelings and laughs and cries easily
jealous	A person who feels resentment about what other people have
self-pitying	A person who feels sorry for themselves
defensive	A person who is easily offended and always guards themselves against criticism
grumpy	A person who is bad-tempered and always complaining
anxious	A person who stresses and worries about things
bossy	A person who likes giving people orders and wants things his/her own way
nosey	A person who is overly curious about other people's business

compulsive	A person who has to do things in a certain way and often checks and does things over and over again to make sure they are done exactly right
patient	A person who is able to accept or tolerate delays or problems and is very relaxed about getting things done
courageous	A person who is not afraid to do the right thing, even if it is dangerous to them
intellectual	A person who thinks a lot about the deeper meaning of things and likes to analyze things
articulate	A person who speaks fluently and clearly, and who can express their ideas well
creative	A person who has good imagination or original ideas
clever	A person who is quick to understand and learn, and who can figure things out quickly
intense	A person who is very serious and expresses strong feelings
independent	A person who is able to think and act without being influenced by others
self-critical	A person who holds himself/herself responsible for any failures, always questioning if they did the right thing or not
curious	A person who is eager to learn about or experience new things
ethical	A person who is careful to do things that are morally right to do
traditional	A person who likes to do things the way they have always been done and accepted in the past
shallow	A person who is concerned only about silly or inconsequential things; superficial
ignorant	A person who doesn't know anything, and is also usually unaware of that
condescending	A person who thinks he/she is better than others and puts other people down
open-minded	A person who is willing to try new things or to hear and consider new ideas
empathetic	A person who is able to understand and share the feelings of others
easygoing	A person who is relaxed, tolerant, and not prone to rigid rules or bouts of temper
determined	A person who is able to make firm decisions and is resolved not to change them
mean	A person who is unkind, inconsiderate, and doesn't share things
SOCIAL EVALUATIONS	
competent	A person who is efficient and capable to do things in general
leader-like	A person who can take charge and help a group accomplish a goal
trustworthy	A person who can be relied on as honest and truthful
charismatic	A person who is interesting and likeable because they have a charming personality
beautiful	A person who looks appealing and physically attractive
weird	A person who does strange or bizarre things
criminal	A person who looks like they could commit a crime
PHYSICAL APPEARANCES	

baby-faced	A person who has facial features resembling a baby
strong	A person who is physically vigorous and is able to exert great bodily or muscular power
youthful	A person who looks young
atypical	The structure, texture, shape or other aspects of the appearance of the face is unusual or rare
EMOTIONS	
happy	A person who is usually cheerful
angry	A person who is usually angry
DEMOGRAPHIC CHARACTERISTICS	
White	A person whose face looks like they are Caucasian
well educated	A person who has completed a high level of education, such as bachelor's, master's and doctorate degrees
INCOME	N/A
homosexual	A person who is sexually attracted to people of his/her own sex
HEALTH	
healthy	A person who is in good health
Autism	A person who has autism spectrum disorder--a developmental disorder characterized by troubles with social interaction and communication, and by restricted and repetitive behavior
CURSE WORDS	
idiot	A person who is stupid
loser	A person who fails frequently or is generally unsuccessful in life

Note: Definitions of the traits were obtained from Google dictionary, with necessary modifications to make the definitions easy to understand and fit the context of describing a person.

4.6 References

1. Oosterhof, N. N. & Todorov, A. The functional basis of face evaluation. *Proc. Natl. Acad. Sci.* **105**, 11087–11092 (2008).

2. Sutherland, C. A. M. *et al.* Facial First Impressions Across Culture: Data-Driven Modeling of Chinese and British Perceivers' Unconstrained Facial Impressions. *Pers. Soc. Psychol. Bull.* **44**, 521–537 (2018).

3. Rule, N. O., Ambady, N. & Hallett, K. C. Female sexual orientation is perceived accurately, rapidly, and automatically from the face and its features. *J. Exp. Soc. Psychol.* **45**, 1245–1251 (2009).

4. Na, J. & Kitayama, S. Spontaneous Trait Inference Is Culture-Specific: Behavioral and Neural Evidence. *Psychol. Sci.* **22**, 1025–1032 (2011).

5. Engell, A. D., Haxby, J. V. & Todorov, A. Implicit Trustworthiness Decisions: Automatic Coding of Face Properties in the Human Amygdala. *J. Cogn. Neurosci.* **19**, 1508–1519 (2007).

6. Cogsdill, E. J., Todorov, A. T., Spelke, E. S. & Banaji, M. R. Inferring Character From Faces: A Developmental Study. *Psychol. Sci.* **25**, 1132–1139 (2014).

7. Adams Jr, R. B., Nelson, A. J., Soto, J. A., Hess, U. & Kleck, R. E. Emotion in the neutral face: A mechanism for impression formation? *Cogn. Emot.* **26**, 431–441 (2012).

8. Ewing, L., Caulfield, F., Read, A. & Rhodes, G. Perceived trustworthiness of faces drives trust behaviour in children. *Dev. Sci.* **18**, 327–334 (2015).

9. Porter, S., ten Brinke, L. & Gustaw, C. Dangerous decisions: the impact of first impressions of trustworthiness on the evaluation of legal evidence and defendant culpability. *Psychol. Crime Law* **16**, 477–491 (2010).
10. Wilson, J. P. & Rule, N. O. Hypothetical Sentencing Decisions Are Associated With Actual Capital Punishment Outcomes: The Role of Facial Trustworthiness. *Soc. Psychol. Personal. Sci.* **7**, 331–338 (2016).
11. Lin, C., Adolphs, R. & Alvarez, R. M. Cultural effects on the association between election outcomes and face-based trait inferences. *PLOS ONE* **12**, e0180837 (2017).
12. Lin, C., Adolphs, R. & Alvarez, R. M. Inferring Whether Officials Are Corruptible From Looking At Their Faces. *Psychol. Sci.* **29**, 1807–1823 (2018).
13. Todorov, A. Inferences of Competence from Faces Predict Election Outcomes. *Science* **308**, 1623–1626 (2005).
14. Blair, I. V., Judd, C. M. & Chapleau, K. M. The Influence of Afrocentric Facial Features in Criminal Sentencing. *Psychol. Sci.* **15**, 674–679 (2004).
15. Wilson, J. P. & Rule, N. O. Facial Trustworthiness Predicts Extreme Criminal-Sentencing Outcomes. *Psychol. Sci.* **26**, 1325–1331 (2015).
16. Rule, N. O. & Ambady, N. Democrats and Republicans Can Be Differentiated from Their Faces. *PLOS ONE* **5**, e8733 (2010).
17. Todorov, A. *Face value: The irresistible influence of first impressions*. (Princeton University Press, 2017).
18. Porter, S., England, L., Juodis, M., ten Brinke, L. & Wilson, K. Is the face a window to the soul? Investigation of the accuracy of intuitive judgments of the

trustworthiness of human faces. *Can. J. Behav. Sci. Rev. Can. Sci. Comport.* **40**, 171–177 (2008).

19. Penton-Voak, I. S., Pound, N., Little, A. C. & Perrett, D. I. Personality Judgments from Natural and Composite Facial Images: More Evidence For A "Kernel Of Truth" In Social Perception. *Soc. Cogn.* **24**, 607–640 (2006).
20. Rule, N. O., Garrett, J. V. & Ambady, N. On the Perception of Religious Group Membership from Faces. *PLOS ONE* **5**, e14241 (2010).
21. Rule, N. O. *et al.* Face value: Amygdala response reflects the validity of first impressions. *NeuroImage* **54**, 734–741 (2011).
22. Olivola, C. Y. & Todorov, A. Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *J. Exp. Soc. Psychol.* **46**, 315–324 (2010).
23. Gordon W. Allport & Henry S. Odbert. Trait-names: A psycho-lexical study. *Psychol. Monogr.* **47**, i (1936).
24. Saucier, G., Hampson, S. E. & Goldberg, L. R. Cross-language studies of lexical personality factors. in *Advances in personality psychology* 1–36 (2000).
25. Saucier, G. & Goldberg, L. R. Evidence for the Big Five in analyses of familiar English personality adjectives. *Eur. J. Personal.* **10**, 61–77 (1996).
26. Goldberg, L. R. An Alternative 'Description of Personality': The Big-Five Factor Structure. *J. Pers. Soc. Psychol.* **59**, 1216 (1990).
27. Saucier, G. & Srivastava, S. What makes a good structural model of personality? Evaluating the Big Five and alternatives. in *Handbook of personality and social psychology* **3**, 283–305 (2015).

28. DeYoung, C. G. *et al.* Testing Predictions From Personality Neuroscience: Brain Structure and the Big Five. *Psychol. Sci.* **21**, 820–828 (2010).

29. Sampaio, A., Soares, J. M., Coutinho, J., Sousa, N. & Gonçalves, Ó. F. The Big Five default brain: functional evidence. *Brain Struct. Funct.* **219**, 1913–1922 (2014).

30. Xu, J. & Potenza, M. N. White Matter Integrity and Five-Factor Personality Measures in Healthy Adults. *NeuroImage* **59**, 800–807 (2012).

31. Privado, J., Román, F. J., Saénz-Urturi, C., Burgaleta, M. & Colom, R. Gray and white matter correlates of the Big Five personality traits. *Neuroscience* **349**, 174–184 (2017).

32. McCrae, R. R. & Costa Jr, Paul T. Toward a new generation of personality theories: Theoretical contexts for the Five-Factor Model. in *The Five-Factor Model of personality: Theoretical perspectives* 51–87 (Guilford Press, 1996).

33. Digman, J. M. Personality Structure: Emergence of the Five-Factor Model. *Annu. Rev. Psychol.* **41**, 417–440 (1990).

34. Dubois, J., Galdi, P., Han, Y., Paul, L. K. & Adolphs, R. Resting-state functional brain connectivity best predicts the personality dimension of openness to experience. *bioRxiv* (2018). doi:10.1101/215129

35. Allen, T. A. & DeYoung, C. G. *Personality Neuroscience and the Five Factor Model*. **1**, (Oxford University Press, 2016).

36. Adelstein, J. S. *et al.* Personality Is Reflected in the Brain's Intrinsic Functional Architecture. *PLOS ONE* **6**, e27633 (2011).

37. Todorov, A., Olivola, C. Y., Dotsch, R. & Mende-Siedlecki, P. Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annu. Rev. Psychol.* **66**, 519–545 (2015).

38. Sutherland, C. A. M. *et al.* Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition* **127**, 105–118 (2013).

39. Walker, M. & Vetter, T. Changing the personality of a face: Perceived Big Two and Big Five personality factors modeled in real photographs. *J. Pers. Soc. Psychol.* **110**, 609–624 (2016).

40. Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. Evaluating the Use of Exploratory Factor Analysis in Psychological Research. 28

41. Hehman, E., Xie, S. Y., Ofosu, E. K. & Nespoli, G. Assessing the point at which averages are stable: A tool illustrated in the context of person perception. (2018). doi:10.31234/osf.io/2n6jq

42. DeBruine, L. & Jones, B. Face Research Lab London Set. (2017). doi:10.6084/m9.figshare.5047666.v3

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

44. Chelnokova, O. *et al.* Rewards of beauty: the opioid system mediates social motivation in humans. *Mol. Psychiatry* **19**, 746–747 (2014).

45. King, D. E. Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* 41755–1758 (2009).

46. Michael Quinn Patton. *Qualitative Research & Evaluation Methods*. (SAGE Publications, 2002).

47. Hehman, E., Sutherland, C. A. M., Flake, J. K. & Slepian, M. L. The unique contributions of perceiver and target characteristics in person perception. *J. Pers. Soc. Psychol.* **113**, 513–529 (2017).

48. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* 135–146 (2017).

49. Zwick, W. R. & Velicer, W. F. Comparison of five rules for determining the number of components to retain. *Psychol. Bull.* **99**, 432–442 (1986).

50. Çokluk, Ö. & Koçak, D. Using Horn's Parallel Analysis Method in Exploratory Factor Analysis for Determining the Number of Factors. *Educ. Sci. Theory Pract.* **16**, 537–551 (2016).

51. Pearson, R., Mundfrom, D. & Piccone, A. A Comparison of Ten Methods for Determining the Number of Factors in Exploratory Factor Analysis. **39**, 15 (2013).

52. Stolier, R. M., Hehman, E. & Freeman, J. B. A common trait space across social cognition. doi:10.31234/osf.io/5na8m

53. Vernon, R. J. W., Sutherland, C. A. M., Young, A. W. & Hartley, T. Modeling first impressions from highly variable facial images. *Proc. Natl. Acad. Sci.* **111**, E3353–E3361 (2014).

54. Fiske, S. T., Cuddy, A. J. C. & Glick, P. Universal dimensions of social cognition: warmth and competence. *Trends Cogn. Sci.* **11**, 77–83 (2007).

55. Stolier, R. M., Hehman, E., Keller, M. D., Walker, M. & Freeman, J. B. The conceptual structure of face impressions. *Proc. Natl. Acad. Sci.* **115**, 9210–9215

56. Markant, J. & Scott, L. S. Attention and Perceptual Learning Interact in the Development of the Other-Race Effect. *Curr. Dir. Psychol. Sci.* **27**, 163–169 (2018).

57. Freeman, J. B. & Johnson, K. L. More Than Meets the Eye: Split-Second Social Perception. *Trends Cogn. Sci.* **20**, 362–374 (2016).

58. Brooks, J. A. & Freeman, J. B. Conceptual knowledge predicts the representational structure of facial emotion perception. *Nat. Hum. Behav.* **2**, 581–591 (2018).

59. Firestone, C. & Scholl, B. J. Cognition does not affect perception: Evaluating the evidence for 'top-down' effects. *Behav. Brain Sci.* **39**, e229 (2016).

60. Bruner, J. S. & Tagiuri, R. The perception of people. in *Handbook of social psychology* **2**, (Addison Wesley, 1954).

61. Kuusinen, J. Factorial invariance of personality ratings. *Scand. J. Psychol.* **10**, 33–44 (1969).

62. Mulaik, S. A. Are personality factors raters' comceptual factors? *J. Consult. Psychol.* **28**, 506–511 (1964).

63. Said, C. P., Sebe, N. & Todorov, A. *BRIEF REPORTS Structural Resemblance to Emotional Expressions Predicts Evaluation of Emotionally Neutral Faces*. (2009).

64. Stolier, R. M., Hehman, E. & Freeman, J. B. Conceptual structure shapes common trait space across social cognition. (2018). doi:10.31234/osf.io/5na8m

65. Brooks, J. A. & Freeman, J. B. Neuroimaging of person perception: A social-visual interface. *Neurosci. Lett.* **693**, 40–43 (2019).

66. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **86**, 420–428 (1979).

67. Lorenzo-Seva, U. & ten Berge, J. M. F. Tucker's Congruence Coefficient as a Meaningful Index of Factor Similarity. *Methodology* **2**, 57–64 (2006).

Chapter 5

General Discussion

5.1 Summary of Findings

How do people make trait inferences from faces? Across three projects we showed that given a context, people made attributions of a small number of traits that were relevant to the context from faces (Chapter 2 & 3) and the context-relevance of a traits was modified by culture (Chapter 2); these trait attributions from faces were influenced by changes in certain facial metrics even for the same facial identity (Chapter 3) and they had consequences for important social outcomes in the real world (Chapter 2 & 3). While people made attributions of a large number of traits when no specific context was given, these attributions could be largely represented in a much lower-dimensional space (Chapter 4).

Specifically, we found that trait attributions from faces were associated with both positive and negative social outcomes in two contexts (electoral success in Chapter 2: Fig. 2; political corruption in Chapter 3: Tables 1 & 2). The association between past election outcomes (which were determined by real-world voters) and trait attributions from politician faces (which were made by participants in laboratory studies, unfamiliar with both the identity of the faces and any information about election outcomes) was stronger when participants made the attributions more rapidly (Chapter 2: Fig. 3). These findings suggest that rapidly formed impressions from political candidate faces might have influenced voting decisions of real-world voters. While attributions of competence were associated with election outcomes in both the United States and South Korea (Chapter 2: Table 2),

attributions of open-mindedness and threat were only associated with election outcomes in South Korea (Chapter 2: Tables 3 & 4). These findings suggest that given the same context (real-world voters deciding which political candidates to vote for), culture might be a crucial factor in determining the relevant trait attributions that people make from faces.

As for the negative context, we showed that how corruptible a politician looked (as judged by participants who knew nothing about the facial identities except that they were politicians) was associated with whether the politician was convicted of corruption and found to have violated the law (Chapter 3: Tables 1 & 2). This association was found specifically for trait attributions that were relevant to the context of political corruption/campaign-law violation (e.g., corruptibility, trustworthiness, selfishness), but not for other trait attributions (e.g., competence, masculinity; Chapter 3: Fig. 3). These findings suggest the possibility that the association between corruption records and perceived corruptibility from politician's faces could result, at least in part, from the biases of prosecutors, judges, and juries by the impressions they formed from the politicians' faces. By comparing the attribution of corruptibility from faces with multiple facial metrics, and by digitally manipulating these facial metrics of the facial stimuli, we demonstrated that facial width (relative to facial height) was one of the underlying facial features that determined the attribution of corruptibility from politician faces (Chapter 3: Fig. 5, Fig. S8).

When attributions from faces were made without any context on an inclusive, representative, and non-redundant set of traits (Chapter 4: Fig. 1), we found that these trait attributions fell along four dimensions: critical/condescending, leadership/competence, female-stereotype, and youth-stereotype (Chapter 4: Fig. 3). These four dimensions were largely reproduced across different cultures, languages, somewhat different experimental

procedures and trait-sets, and even in single subjects (Chapter 4: Fig. 5 & 8). These four dimensions were not simply semantic dimensions of the trait-set (Chapter 4: Fig. 7). They corresponded to the dimensions found in other literatures of person perception and social cognition (Chapter 4: Fig. 6). Our data also revealed that people were able to make reliable and consensual attributions of a large number of traits from faces (Chapter 4: Fig. 2).

5.2 Limitations

This thesis, and the entire literature on trait attributions from faces, has focused on attributions that are artificial in several aspects, raising critical questions regarding the extent to which they characterize attributions people make in real life. First, laboratory attributions (i.e., trait attributions from faces collected in laboratory or online studies) are typically made very rapidly (typically ranging from 30 milliseconds to 2 seconds) and without any incentives (participants' judgments have no consequence), whereas those in real life often take minutes, hours, or even days, and are consequential (one reason why people would take time to make these attributions; e.g., deciding whether to date or employ someone given their photos and profiles). Little is known how trait attributions from faces would change when the depth of processing increases (e.g., when participants are given a longer time to look at the faces and a higher incentive to make accurate attributions), which is an important future direction.

Second, and relatedly, laboratory attributions are generally about unfamiliar faces (i.e., first impressions from faces), whereas those in real life also concern people with whom we are familiar—or with whom we become familiar in the course of making lengthy series of attributions about them. For example, an individual might encounter a political candidate's

campaign photos on social media multiple times before deciding whether to look up detail information of the candidate. Little is known how second impressions, third impressions, etc. are formed and how they might be related to first impressions.

Third, and also relatedly, laboratory attributions are generally made to static images of isolated faces, omitting the dynamic changes in faces that are accompanied by eye movements, facial expressions, and body postures in real life. Finally, even more broadly, laboratory attributions omit rich contextual information as compared to those in real life: participants do not find out what other people think about the individual whose face they are viewing, read no biography of the individual, and do not see the individual behaves or speaks in the context.

5.3 Future Directions

Related to the limitations mentioned above, an important future direction is to study trait attributions from faces using experimental designs that are more similar to the situations in real life. For example, by extending the exposure time of the face and introducing rating-dependent rewards or punishments, we can understand how people make trait attributions from faces when they have a longer time to look at the face (e.g., first, second, third impressions) and when their evaluation is consequential.

This thesis has also focused on behavioral data only, leaving open important questions about the underlying biological mechanisms. An important future direction is to investigate the brain mechanisms that might give us insight into the psychological stages of making trait attributions from faces. What are the facial features the brain uses to evaluate a face (across various contexts)? Where in the brain are those evaluations computed? Are those evaluation-

related facial features tracked by distinct networks of the brain? To answer the first question, recall that trait attributions from faces are likely to rely on holistic processing of the face as a whole, instead of differentiable processing of individual features of the face (Chapter 1). The four dimensions found in Chapter 4 that represented the attributions of a large and comprehensive set of traits are presumably the best candidates for the holistic features that the brain might use to compute the value of a face. Therefore, future research should investigate whether subjective ratings of how critical, leader-like, feminine, and youthful (i.e., the four dimensions found in Chapter 4) a face looks predict the subjective value of a face (e.g., the value elicited by asking "how much are you willing to use the face as your own profile photo?"). To answer the second question (i.e., from which region of the brain could we decode the value of a face in various contexts), a potential approach is to train a linear classifier on patterns of fMRI response to categorize faces of high and low values in different contexts. To answer the third question (i.e., do distinct patterns of voxel activity in the value-coding region represent each of the four holistic features), a potential approach is to test for each pair of the four holistic features, whether the classifier trained on the value of one feature could predict the value of another feature (the prediction should not be significant if each feature correspond to a distinct pattern).

An alternative approach could be to disregard the four psychological dimensions found in Chapter 4 altogether, and ask about the relative weight of each individual trait attribution in constructing the subjective value of a face (across different contexts). Perhaps the brain's representation of the value of a face is inherently more flexible or higher dimensional—that is, this representation can be used both to derive the psychological space generated when people are explicitly asked to attribute specific traits (as in Chapter 4) and also to derive the

value of a face as computed on weighted trait attributions. The question, “how do people represent traits in faces?” may thus not have a single answer.