Universität Basel
Fakultät für Psychologie

Refined Reverse Correlation:

A Technique for Investigating the Power of Faces

Inauguraldissertation zur Erlangung der Würde eines Doktors der Philosophie vorgelegt der Fakultät für Psychologie der Universität Basel von

Matthias David Keller

aus Oberthal (BE), Schweiz

Basel, 2019

Universität Basel

Fakultät für Psychologie

Genehmigt von der Fakultät für Psychologie auf Antrag von

Prof. Dr. Rainer Greifeneder
Prof. Dr. Roland Imhoff

Datum des Doktoratsexamen: 12.09.2019

Prof. Dr. Alexander Grob
Dekan der Fakultät für Psychologie

Universität Basel

Fakultät für Psychologie

Erklärung zur wissenschaftlichen Lauterkeit

Ich erkläre hiermit, dass die vorliegende Arbeit ohne die Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel selbstständig verfasst habe. Zu Hilfe genommene Quellen sind als solche gekennzeichnet. Die veröffentlichten oder zur Veröffentlichung in Zeitschriften eingereichten Manuskripte wurden in Zusammenarbeit mit den Koautoren erstellt und von keinem der Beteiligten an anderer Stelle publiziert, zur Publikation eingereicht, oder einer anderen Prüfungsbehörde als Qualifikationsarbeit vorgelegt. Es handelt sich dabei um folgende Manuskripte:

- (I) Rudert, S.C., **Keller, M. D.**, Hales, A. H., Walker, M., & Greifeneder, R. (2019). *Who gets ostracized? A personality perspective on risk and protective factors of ostracism.* Manuscript in revision for publication.
- (II) **Keller, M. D.**, Reutner, L., Greifeneder, R., & Walker, M. (2019). *Faces evoking emotions stereotypically triggered by groups: Developing an advanced reverse correlation technique.* Manuscript under review.
- (III) Walker, M. & **Keller, M. D.** (2019). Beyond attractiveness: A multimethod approach to study enhancement in self-recognition on the Big Two personality dimensions. *Journal of Personality and Social Psychology*. Advance online publication. doi:10.1037/pspa0000157
- (IV) Stolier, R. M., Hehman, E., **Keller, M. D.**, Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 9210-9215. doi:10.1073/pnas.1807222115

Basel, 11.06.2019

Matthias David Keller

Acknowledgements

First and foremost, I would like to thank Mirella Walker who was my mentor and supervisor, already throughout my time as a research assistant, and afterwards throughout my whole PhD. Without Mirella none of these projects would have become reality. I would also like to sincerely thank Leonie Reutner and Rainer Greifeneder, who assisted me at different times in my dissertation as supervisors and continuously provided me with support and advice. I am also very grateful to Roland Imhoff for fruitful discussions at conferences and for agreeing to be part of my PhD committee.

I would also like to express my special thanks of gratitude to all my collaborators for all their hard work and the ongoing fruitful exchange throughout the projects and beyond. Furthermore, I could always rely on a team that provided me with feedback and good advice, for which I am very thankful. In particular, I would like to thank Mariela Jaffé for an enormously helpful and pleasant office partnership through the last four years, for her feedback on my thesis, for listening to my concerns, and much more. I would also like to thank my research assistants, Antonin Tröndle, Rebecca Götsch, and Mirjam Thali, for their tremendous work throughout the years. I would also like to thank Caroline Tremble for her help in formulating my thoughts in proper English, which was not necessarily the part of my work that was the most enjoyable, especially at the beginning of my PhD.

Special thanks go to my parents, who have always supported me, always had time and advice for me and taught me so many important things in life. Last but not least, I would like to thank Katja Schönfeld for everything she has done and is doing with and for me.

Table of Contents

Abstract

People effortlessly and rapidly form a first impression of an individual's personality based on their facial appearance. Forming an impression based on facial cues can have real world implications, for example, for the outcome of elections, courtroom decisions or work-place interviews. Research using traditional methods has, however, failed to identify the facial features that are related to specific personality traits in a reliable and valid way. This challenge can be overcome using a reverse correlation method. Here I present a refinement of the traditional reverse correlation image classification technique. Over the course of four projects I highlight the different possibilities that the refined technique offers. In the first project I will present how the technique was used to extract the facial prototype of someone that is likely to be ostracized. In the second project, I show how we extracted prototypes that evoke different emotions, applied them to real facial photographs and set the different prototypes in relation with each other. The third project offers insights into how the technique was used to investigate self-perception without any external standard of comparison except the participants' own face. Finally, I present a fourth project where the technique was used to investigate whether the belief about how two personality traits co-occur on a conceptual level is reflected in the facial characteristics that are used to form an impression from faces. The here presented refined technique adds to the traditional reverse correlation technique in that internal representations can be visualized without visible artifacts, that the extracted prototypes can be applied to real photographs, and set in relation with each other. The discussion focuses on the reliability and validity of the method and presents future research possibilities.

Introduction

When asked why he had supported Warren G. Harding's political campaign to become the 29th President of the United States, Harry Daugherty responded with "he looked like a president" (The United States Governement, n.d.). This quote can be understood either in the figurative sense, that Daugherty believed that Harding would make a good president, or in the literal sense, that Daugherty believed that Harding actually *looked* like a president. Literature in the domain of face perception points out that the literal sense is not as farfetched as it might seem at first. People do infer personality from faces and even act upon these ascriptions. A physiognomist (someone who studies the outer appearance of a person to gain knowledge about the character of that person) once wrote that the forehead of Harding "indicates broad-mindedness, and intellectual powers" (LeBarr, 1922, pp. 139), personality traits that can be assumed to be important in a leader. Thus, it seems to be the case that there was something in Harding's face that made him look like a president and that it was not only Daugherty who *saw* this. Furthermore, if we follow the reasoning of the literal interpretation of Daugherty's quote, we might conclude that he also assumed that other people might have this impression and, if one goes one step further, may even act in line with this impression by voting for Harding.

However, not all elected presidents look exactly alike. But it might be that the likelihood of being elected as a politician rises if specific facial characteristics, such as a forehead that indicates broad-mindedness, are prominent. Thus, the question arises as to whether it is a more complex facial structure that leads to the perception of possessing 'intellectual powers', or even to the perception of being electable. What would such a face look like? Can we even put into words what makes a face electable? Or might it be more intuitive to visualize it?

In this dissertation, I present a refined reverse correlation technique, that enables the visualization of prototypes in a highly realistic manner in multiple faces. This refined

technique combines the image classification task from the traditional reverse correlation technique (Dotsch, Wigboldus, Langner, & van Knippenberg, 2008; Kontsevich & Tyler, 2004; Mangini & Biederman, 2004) with statistical face modeling (Paysan, Knothe, Amberg, Romdhani, & Vetter, 2009) and up-to-date computer graphics (Walker & Vetter, 2016). The refined reverse correlation technique presented here is a very powerful tool to investigate face perception in many different domains and from many different angles. By using the image classification task, the assignment is very intuitive for participants, which enables the visualization of otherwise hidden characteristics and renders the technique less prone to social desirability issues. Furthermore, by incorporating the technique into a statistical face space, the multiple relationships between different prototypes can be explored and defined. Finally, by using up-to-date computer graphics this technique enables one to apply the extracted prototypes to multiple faces with realistic results. Consequently, more complex study designs are possible, enabling a higher generalizability of the results.

The dissertation is structured as follows: I will first outline the importance of research on face perception and why we need sophisticated methods in this particular domain of research. Next, I will present a method that enables the extraction of facial prototypes in a data-driven manner, namely reverse correlation. I will then discuss in more detail the classical reverse correlation technique that uses the image classification task and how we combined the task and statistical face modeling in order to simultaneously profit from the advantages that the image classification task and statistical face modeling offer. Four different projects will be presented in which the benefits of our novel technique, hereafter referred to as the refined reverse correlation technique, will be apparent. Lastly, I will discuss the reliability and validity of the presented technique and highlight its potential for further research.

The Importance of Faces

The physiognomic account holds that personality can validly be observed from faces. This belief has a long history and physiognomy reached its sad climax as a so-called science

in World War II (Gray, 2004). However, the lay belief that certain personality traits are visible in faces still persists (Suzuki, Tsukamoto, & Takahashi, 2019) and some research even points out that faces may actually contain valid information that can be extracted via machine learning. Wu and Zhang (2016), for example, developed an algorithm that was supposedly able to distinguish between faces that belong to convicted criminals and those that belong to non-criminal individuals. In more recent work, Wang and Kosinski (2018) developed an algorithm that is supposedly able to reliably distinguish between faces of self-identified heterosexual and homosexual men and women. However, it is more likely that the algorithm only identifies differences in socieconomic status in the former and stereotypes in the latter (see also Agüera y Arcas, Todorov, & Mitchell, 2018), instead of actually distinguishing between two supposedly different groups of people based on their facial features. If objective machine learning approaches do not seem to be able to distinguish between people regarding specific group affiliation, then how should people be able to make ecologically valid assumptions about a person based on a face? Supporting this assumption, there is a large body of research suggesting that the ecological validity of face based personality aascritipions is rather negligible, especially when controlling for gender, ethnecity, and age (Olivola, Funk, & Todorov, 2014; Pound, Penton-Voak, & Brown, 2007; Shevlin, Walker, Davies, Banyard, & Lewis, 2003; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015).

Irrespective of whether trait ascriptions based on faces have ecological validity or not, faces are widely used as a cue to infer something about the other person's personality (e.g., Bruce & Young, 1986; Willis & Todorov, 2006). Moreover, these judgements have real world implications and influence decisions and behavior, for example, in voting decisions (Ballew & Todorov, 2007; Little, Burriss, Jones, & Roberts, 2007; Olivola & Todorov, 2010; Todorov, Mandisodza, Goren, & Hall, 2005) or in criminal sentencing (Porter, ten Brinke, & Gustaw, 2010; Sigall & Ostrove, 1975; Zebrowitz & Mcdonaldt, 1991). Research in the domain of antecedents and consequences of face perception has therefore attracted a lot of

attention, especially in the field of social psychology. However, in order to investigate the antecedents of perceived personality or stereotypes in faces, they first need to be identified.

How to Identify Facial Characteristics of Prototypes

To identify facial characteristics related to a personality trait, researchers could simply ask people to name the facial cues they associate with that trait. For example, to identify the facial characteristics of an electable face, researchers could ask people for the facial characteristics they associate with a political leader or a person they would potentially vote for. However, people seem rather unable to verbalize the specific cues in faces that they rely on when making a snap judgement about someone, especially because of the complexity of face perception (Todorov, Loehr, & Oosterhof, 2010). Moreover, faces are processed in a holistic manner, meaning that the same part of a face (e.g., curved eyebrows) results in a different ascription when presented in combination with other facial features (e.g., mouth tips are either pointing down- or upwards; Rossion, 2013).

Thus, in order to investigate and understand the role of facial features that lead to the ascription of a certain personality or a certain stereotype, a methodological approach is required that overcomes these challenges. As people generally agree on what a face looks like that signals a specific personality trait (e.g., Todorov, Said, Engell, & Oosterhof, 2008; Zebrowitz & Montepare, 2008), the specific characteristics that people agree on might be visualized. This challenge can be addressed with reverse correlation techniques (Dotsch, Wigboldus, Langner, & van Knippenberg, 2008; Oosterhof & Todorov, 2009; Walker & Vetter, 2009, 2016; for an overview, see Todorov et al., 2011).

Traditional Reverse Correlation Technique

The term *reverse correlation* refers to techniques in which the variation in stimulus attributes is not meaningfully manipulated, but is random. Instead of establishing a correlation between manipulated attributes and participants' responses, reverse correlation methods use the correlation between a fixed response variable and random stimuli in order to model the

attributes of the stimuli that caused the observed choice pattern of participants (Todorov et al., 2011).

Mangini and Biederman (2004) and Kontsevich and Tyler (2004) were the first to use a reverse correlation approach[1] in order to model the information that mediates face classification into a specific category (e.g., sad vs. happy). What renders the task very intuitive for participants and thus enables the extraction of otherwise covered information is the use of randomly distorted images that need to be classified into a specific category. An exemplary reverse correlation image classification study consists of three ingredients: first, a so-called base face, which is used as the basis for all stimuli throughout the study; second, random noise patterns, which will be superimposed onto the base face; and, third, the categories into which images should be classified.

Dotsch and colleagues (2008) slightly altered the initial task of classifying a random face stimulus into one of multiple categories to a task where one of two opposing random face stimuli needed to be classified into a specific category. To create one trial, a random noise pattern is both added to and subtracted from the base face. Thus, for every trial, two opposing versions of the base face are created. During the study, participants are repeatedly presented with these two versions stemming from the same random noise pattern side by side, and then asked to decide which of the two versions resembles their internal representation of the category of interest (e.g., choose the more Moroccan-looking face; Dotsch et al., 2008). By averaging all noise patterns selected by a specific individual or of all participants within the same condition, an average noise pattern can be calculated that is no longer random but contains meaningful information, in that it reveals the averaged information participants used in the task to make their decisions. The application of the average noise pattern to the base face reveals the internal representation of either a specific individual or of all participants

[1] Originally, the method was developed in the domain of auditory cognition (Ahumada, 2002; Solomon, 2002).

within the same condition (e.g., the prototypical Moroccan). Compared to openly asking participants about their stereotypes, using the reverse correlation image classification task renders the task very intuitive for participants and allows the extraction of (even subtle) facial characteristics that would have otherwise remained covert. Additionally, due to the implicit nature of the task it is less prone to social desirability concerns.

This method is widely used in social psychological research and has undeniably proved to be very useful for visualizing, for example, personality dimensions (Dotsch & Todorov, 2012; Oliveira, Garcia-Marques, Dotsch, & Garcia-Marques, 2019), identity (e.g., Mangini & Biederman, 2004; Young, Ratner, & Fazio, 2014), or groups (e.g., Brown-Iannuzzi, Dotsch, Cooley, & Payne, 2016; Dotsch et al., 2008; Imhoff, Dotsch, Bianchi, Banse, & Wigboldus, 2011; Imhoff, Woelki, Hanke, & Dotsch, 2013). However, the traditional technique faces some challenges, especially if the research aim goes beyond the mere visualization of prototypes. All the challenges that I discuss here result from the static nature of random noise patterns that are used to distort the base face. The resulting classification image, thus, is also static in nature because it is dependent on the base face that has been used in the classification task. This means that, first, the resulting noise pattern can only be meaningfully applied onto the same base face that has been used during the image classification task. Second, the resulting average noise pattern is a blurry greyish pixel-pattern. Applying this average noise pattern onto the base face also results in a rather blurry black and white image. Using these resulting prototypes as realistic stimuli in future research is therefore not possible. Both these issues are problematic if one is aiming for high generalizability and to treat stimuli as a random effect in a mixed-effects model (Judd, Westfall, & Kenny, 2012, 2017). In such a study design, multiple stimuli would be needed that could be classified as realistic looking faces. Third, different classification images as well as the underlying random noise patterns cannot be related to each other in a meaningful way. All the above-mentioned challenges can be solved by the substitution of the random noise

patterns with random vectors stemming from a statistical face space. Overcoming these boundaries enables a multitude of interesting possibilities, such as relating different extracted prototypes directly to each other, or using the same prototype to manipulate different faces.

Operating with Random Vectors

In order to understand why random vectors instead of random noise patterns can be used, I will introduce the idea of a face space. The assumption of a face space (Valentine, 1991) holds that every face can be located as a point in a multidimensional space. This means that if two faces are perceptually similar to each other, these two faces are also located in close proximity to each other in this multidimensional space. A third face that differs from the first two faces among many (or all) dimensions in the face space would conversely be located farther away from the first two faces. Based on these assumptions, Paysan, Knothe, Amberg, Romdhani, and Vetter (2009) created a statistical face space; the Basel Face Model (BFM; Paysan et al., 2009). This face space is built with 100 male and 100 female 3D scans. In order to extract the dimensions that best explained the variance between the faces, the 3D scans were mathematically represented by vertices coding for shape and the corresponding color of the vertices. To extract the dimensions that best explained the variance between the different faces, two principal component analyses were conducted, one for shape and one for color. This procedure resulted in a 199-dimensional shape and a 199-dimensional color space. Every face is located both at a specific position within the 199-dimensional shape space and a specific position within the 199-dimensional color space. Thus, every face can be understood as a vector that points from the center of the multidimensional spaces to a specific position. Randomly combining values for the different dimensions within one of the spaces (i.e., random vectors) results in a random location within the multidimensional space. The centers of both the shape and the color space are expressed by the value zero on each of the dimensions within the specific spaces (i.e., zero-vectors). Averaging all the 100 male and 100 female faces results in this zero-vector position, by definition.

By randomly combining values for each dimension the face space is built with, random locations within the face space are established. This random point in the face space can also be understood as a random face stemming from this multidimensional face space. Adding a specific random vector to the average face (i.e., zero-vector face), as well as subtracting this random vector, results in two new faces that are opposing each other mathematically in the face space. The two resulting faces can now be used as one trial for an image classification task.

During the image classification task, the two opposing faces are presented on the same page to the participants. Their task then is to choose the version that better represents their internal representation of the prototype in question. In the final step, simply spoken, all chosen faces can be averaged and the resulting face is thought to represent the internal representation of the prototype. Mathematically speaking, each underlying random vector is included in the averaging process while the degree to which it plays a role in the resulting average vector is identified by participants' choices. For example, if in a specific trial 50 percent of the participants choose one face over the other, this vector does not seem to bear meaningful information and during the averaging process this specific random vector will cancel itself out. But if in a specific trial 90 percent of the participants choose one face over the other, this random vector appears to bear meaningful information and during the averaging process this information will be considered.

Overview of the Dissertation Projects

In total, I will present four different projects where we (i.e., the specific research group) used the refined reverse correlation method to address various research questions within the domain of social psychology. In the first reported project, Rudert, Keller, Hales, Walker, and Greifeneder (2019) visualized the prototype of someone who is likely to be ostracized with the aim to gather insights into the perceived personality of someone who is likely to be ostracized. In the second project, Keller, Reutner, Greifeneder, and Walker (2019)

first, extracted five prototypes of faces that evoke the emotions admiration, envy, pity, disgust, and fear, respectively, in the perceiver, second, went beyond mere visualization of prototypes by showing how we can correlate the different prototypes with each other, and third, presented how the extracted prototypes can be applied to any novel face. In this dissertation's third project, Walker and Keller (2019a) used participants' own face as the base face in the image classification task in order to gain insights into self-perception without using any external standard of comparison. In the fourth project, Stolier, Hehman, Keller, Walker, and Freeman (2018) used the refined reverse correlation technique to dive into the role of individual differences when it comes to the question of what personality looks like in faces.

Dissertation Project 1 – Ostracism

In the first reported research project of this dissertation, Rudert, Keller and colleagues (2019) successfully applied the refined reverse correlation technique to gain insights into the question of what a stereotypical ostracizable person looks like. In this project we aimed to investigate the impact of someone's personality on the likelihood that this person would be ostracized. Together with a longitudinal study and vignette studies we focused on the Big Five personality traits (Costa & McCrae, 1992) and showed that especially agreeableness and conscientiousness are crucial predictors of whether someone will be ostracized or not. In Study 3a we applied the refined reverse correlation technique to visually extract what it is in faces that results in the perception that someone is likely to be ostracized.

We used a morph between the 100 male and 100 female 3D scans from the Basel Face Model (BFM; Blanz & Vetter, 1999) as the base face in this study. Every participant had to indicate, in 200 trials, which of the two versions she or he would rather ostracize. In order to get a measure of reliability we used two different random vector sets and therefore two different sets of faces that participants were presented with. Although participants were presented with different random faces in the two different conditions, the two average vectors point in a very similar direction, which can be concluded by the high weighted correlation

between the two vectors. This finding already supports the method's reliability. In a second step, we aimed to validate the prototype vector and relate it to findings gathered with traditional methodological approaches. We therefore added the extracted prototype vector to the used base face once and subtracted it from the base face once. In the subsequent study, we then presented the two versions to a new set of participants and let them rate the two versions on the Big Five personality traits. The prototypical ostracizable face was perceived as being lower on agreeableness and lower on conscientiousness compared to its counterpart, which is in line with our previous findings reported in that paper. Together, these results offer insights for research on ostracism, by providing a fresh perspective of the perceived personality of the target. Moreover, the results attest to the reliability and validity of this novel method.

The first project showed the method's reliability and pointed out how the extracted vector can be applied onto the same base face that was used during the image classification task in order to visualize and validate the extracted prototype. The next project will go a step further and show how the extracted vectors can be applied to novel faces to strive for a high generalizability of the results. Moreover, I will discuss how different prototypes can be related to each other in order to investigate their similarity or dissimilarity.

Dissertation Project 2 – Emotion

In the second research project of this dissertation, Keller, Greifeneder, Reutner, and Walker (2019) applied the refined reverse correlation technique to extract facial prototypes that evoke specific emotions. Research on group perception holds that the content of stereotypes regarding social groups can be captured by two dimensions, namely warmth and competence (Fiske, Cuddy, Glick, & Xu, 2002). Along with cognitive ascriptions, the Stereotype Content Model (SCM) further identifies specific emotional reactions that are evoked when thinking of either a specific group or an exemplar belonging to this social group. These emotions are admiration, envy, pity, and disgust. Because the SCM spans two dimensions (i.e., warmth and competence; Fiske, Cuddy, Glick, & Xu, 2002) that are similar

to the Big-Two in person perception (i.e., communion and agency; Abele & Wojciszke, 2007) and to the core dimensions in face perception (i.e., trustworthiness and dominance; Oosterhof & Todorov, 2009), the question arises as to whether the emotions that are most prominent in group perception will be evoked from faces in a similar vein and how these emotions are related to each other on the level of individuals.

In the first study of that project, we asked participants to perform an image classification task in which they repeatedly indicated which of two faces elicited more admiration [envy, pity, disgust, fear]. To test the method's reliability we used two different sets of random vectors, as was done in the first reported project (Rudert et al., 2019). In a second step we calculated the resulting prototypes by averaging participants' choices within each of the conditions. The correlation pattern between the different emotion prototypes gives further insights into how the different emotion prototypes are related to each other. The prototype of someone that evokes admiration is highly similar to someone that evokes envy. Likewise, someone that evokes disgust is highly similar to someone that evokes fear, although to a lesser degree.

In order to validate the prototypes, we applied the extracted emotion-prototypes to real photographs from the Basel Face Database (BFD; Walker, Schönborn, Greifeneder, & Vetter, 2018) and asked participants to indicate to what degree these faces evoked the five emotions admiration, envy, pity, disgust, and fear. The results provide strong support that each of the extracted prototypes accurately captures what it is meant to reflect. Thus, for example, faces onto which we applied the admiration prototype vector were perceived as being more likely to evoke admiration than other emotions. Again, we found an admiration-envy, and disgust-fear similarity. Thus, faces that were manipulated to evoke admiration to a similar degree also evoked envy, and vice versa. Two additional studies gave support that this pattern can be observed in non-manipulated faces and on a conceptual level as well. These findings highlight that the admiration-envy, and disgust-fear similarity is not a methodological artifact of the

used technique but is rather inherent in how faces are perceived. Furthermore, this similarity pattern does not appear to be limited to the domain of face perception.

This project demonstrates two important methodological points. First, the extracted prototype vectors can be compared with each other by calculating the correlation between each other. The correlation between the two vectors is an indicator about how much two prototype vectors dissociate between each other and these similarities can be further associated with conceptual ratings. Second, we showed that the extracted prototype vectors can be added onto any face with realistic results. The possibility of adding a prototype vector onto multiple faces enables more complex study designs where faces can be treated as a random effect, which enables findings to be generalized not only across participants but also across facial stimuli.

The first two projects presented how we extract prototypes by using a specific base face, to and from which we add and subtract random vectors. Moreover, the resulting prototype can be applied to any novel face that has previously been located in the statistical face space. However, it is also possible to use different base faces within the same study for the image classification task but still extract a meaningful average vector across all participants. This procedure will be examined in the next project.

Dissertation Project 3 – Self-Perception

In the literature on self-perception there seems to be strong evidence that the *image* one has of oneself is not always perfectly accurate, but often unrealistically positive instead (Alicke & Govorun, 2005; Alicke & Sedikides, 2009). The methodological approaches to measure self-perception in these studies have in common that they compare participants' self-evaluation either with evaluations by others or with their evaluations of others. Thus, there is always an external standard of comparison involved which might already be biased. On the one hand, it might be that individuals' self-evaluations are indeed inflated and the reference

value is accurate. On the other hand, it could also be that the used reference value by itself is deflated and the self-perception is actually accurate.

In an attempt to present a methodological approach to measure self-enhancement with participants' own face as the only indicator, Walker and Keller (2019a) applied the refined reverse correlation approach to the domain of self-perception.

In comparison to the previous reported projects, the used base face was not a morph between different faces but the participant's own face. Thus, in a first step all participants came to our lab where we took photographs of them. Next, we localized each participant's face individually in the multidimensional shape and the multidimensional color space. To create the stimuli for the image classification task, we added random vectors to and subtracted random vectors from every participant's face. The task for the participants was to indicate, in each of the multiple trials, which of the two presented versions was more in line with their true self.

In order to investigate whether people self-enhance on the personality dimensions agency and communion, in an additional study we extracted an agency and a communion vector separately for male and female faces in the same vein as we extracted the ostracizable prototype (Rudert et al., 2019) and the emotion prototypes (Keller et al., 2019). Additionally, we asked participants to indicate their self-esteem with the intention to investigate the role of participants' explicit self-esteem in the domain of self-perception.

We performed two different analyses with the extracted self-perception vectors. In order to investigate whether participants self-enhance on the agency and the communion dimensions, in a first step, we calculated the weighted correlations between each individual random vector and the agency [communion] vector we extracted in the additional study. This gives us an indication of how closely a random vector is associated with the respective personality dimension. In our analysis we then used the absolute amount of the correlation between the random vectors and the personality vectors to predict whether participants would

choose the version that is positively associated with the respective personality trait. Results indicate that without using any standard of comparison, participants self-enhance on the personality traits agency and communion.

In the second analysis we used participants' explicit self-esteem to predict the degree to which participants self-enhance on the Big Two. Therefore, we calculated for each participant their individual self-perception vector. Next, we calculated the weighted correlation between individuals' self-perception vector and the averaged agency [communion] vector. This gives us an indication of the degree of similarity that a participant's self-perception vector shares with the direction of the personality vectors. In our analysis we then used participants' explicit self-esteem to predict the extent to which an individual's self-perception vector is correlated with the agency [communion] vector. Results show that the higher the explicit self-esteem of an individual, the more she or he self-enhances on the Big Two personality dimensions.

This project adds to the literature of self-perception in that it provides a method that enables one to investigate self-enhancement without any external standard of comparison. Furthermore, while not using any external standard of comparison, the project provides evidence that individuals self-enhance on the personality dimensions agency and communion. From the methodological perspective, this project outlines the further possibilities of the refined reverse correlation technique. By using the participants' own faces as the base face in the image classification task, we were able to investigate self-perception without using any standard of comparison. A benefit of this approach is that the same random vector can be applied to different faces while holding the perceptual change in the face constant. If a random vector is strongly associated with, for example, agency, this should result in higher ascribed agency if the vector is added to any face. Moreover, although different faces have been used in the image classification task, the resulting average vectors for each individual

can again be compared with each other and with existing vectors in the multidimensional face space.

So far, we have focused on the consensus people have about what specific personality traits (i.e., Big Two; Walker & Keller, 2019), emotions (Keller et al., 2019) and stereotypes (Rudert et al., 2019) look like in faces, how we can visualize them and how we can relate the resulting prototypes with each other. However, although there is a high consensus about what, for example, a trustworthy face looks like, not every individual agrees on that; there still remains unexplained variance in trait ascriptions. Can our method be used in order to also shed light on this topic?

Dissertation Project 4 – Conceptual/Perceptual

Although there is a high degree of consensus about what specific personality traits look like in faces, there remains unexplained variance (Engell, Haxby, & Todorov, 2007; Hönekopp, 2006). There are several accounts discussing that this dissent in personality trait judgments might be systematic. Research has shown, for example, that a face is evaluated more positively if it resembles that of the judge (Bailenson, Iyengar, Yee, & Collins, 2008). In the here presented project, we show one possibility that sheds light on this unexplained variance, by investigating individuals' lay beliefs about how specific personality traits are related with each other (on a conceptual basis) and whether these structures are also correspondingly reflected in face perception. Put simply, does someone who believes that extroverted people in general are agreeable judge the face of any individual they perceive as extroverted as agreeable as well? Moreover, does this person also use similar facial characteristics when searching for cues that indicate whether someone is extroverted and for someone who is agreeable?

We used the refined reverse correlation technique to answer this question (Stolier et al., 2018). In Study 3 of this project, participants were asked to perform two image classification tasks on two different personality traits from the Big Five (McCrae & Costa,

1997), which resulted in ten possible trait-pair combinations into which participants were randomly allocated. By individually averaging participants' choices for each trait, we extracted two personality vectors for each participant. In the next step, to form an index on how similar the facial characteristics that participants relied on during the image classification task were, we calculated the weighted correlation between the two personality vectors for each participant individually. On the one hand, if a participant uses similar characteristics in faces for both personality traits, the weighted correlation between the two extracted vectors would be high. On the other hand, if a participant uses dissimilar characteristics, this would be reflected in low or even negative correlations between the two personality vectors.

In order to measure how the two traits are believed to co-occur on a conceptual level, we additionally asked participants to indicate how likely it is that someone possesses a personality trait if this person also possesses the other. For example, we asked them to indicate how likely it is that if someone is extroverted this person is also agreeable. The higher someone rates this question, the more likely it is that they believe that two personality traits co-occur on a conceptual level.

In our data we used participants as the unit of analysis. From each participant we used two data points. First, we used the perceptual similarity about how similar the two facial prototypes are. Second, we used the conceptual similarity about how similar the two personality traits are believed to be. As we found a strong correlation between the conceptual and the perceptual similarity we can conclude that the lay belief about how personality traits co-occur impacts how faces are perceived.

Together, these results suggest that our face impressions are also shaped by our conceptual beliefs about how different personality traits are associated with each other. Moreover, these findings are exciting because they show that our refined reverse correlation technique provides insights on an individual level on what facial characteristics individuals

use to make inferences from a face and whether individuals use similar or different facial characteristics to derive a decision.

Discussion

With a methodological focus, this dissertation presents a refined reverse correlation technique that enables the extraction of facial prototypes in a highly intuitive manner with highly realistic results. The technique combines the image classification task (Dotsch et al., 2008; Kontsevich & Tyler, 2004; Mangini & Biederman, 2004) with a statistical face space (Paysan et al., 2009) and uses up-to-date computer graphics (Walker & Vetter, 2016).

With this refined technique we answered research questions about what the prototypical face of someone who is likely to be ostracized looks like and what the associated personality structure of such a prototype is (Rudert et al., 2019), what someone who evokes a specific emotion in perceivers looks like and how these emotion prototypes relate to each other (Keller et al., 2019), whether individuals self-enhance on the Big Two personality dimensions without the use of an external standard of comparison (Walker & Keller, 2019a), and whether similar facial characteristics are used to infer personality from faces when these traits are believed to be correlated in individuals (Stolier et al., 2018).

In the following sections I will first focus on the reliability and validity of the method. Next, I will give an outlook on further possible applications of the refined reverse correlation technique. Finally, I will discuss further possibilities where the technique has either already been applied or might be applied in future research projects.

Reliability and Validity

Throughout the different projects we gathered strong evidence for the reliability of the refined reverse correlation technique. In three of the four reported projects (i.e., Dissertation Project 1 – Ostracism, Dissertation Project 2 – Emotion, and Dissertation Project 3 – Self-Perception) we used two independent vector sets to create the stimulus material for the image classification task. In all three projects and, thus, among eight different prototypes, we found

very high correlations between the averaged vectors derived from the two different vector sets for each of the eight prototypes. These findings indicate that although different stimulus material was used, the two vectors are pointing in a highly similar direction within the multidimensional face space and thus strongly attest the technique's reliability.

We found strong support for the validity of the refined reverse correlation technique in the form of validation studies and in the form of converging evidence between our extracted facial prototypes and conceptual ratings in three projects. Dissertation Project 1 – Ostracism (Rudert et al., 2019) indicates that participants' perception of the extracted ostracizable face is in line with vignette studies as well as with a high powered longitudinal study. In Dissertation Project 2 – Emotion (Keller et al., 2019), there are at least three indicators for the technique's validity. First, in the second study of this project, we successfully validated the different prototypes in that a specific emotion prototype evokes the respective emotion more strongly than the other emotions. Second, the degree to which individuals distinguished between our emotion prototypes can be predicted by the degree to which individuals in general distinguish between these emotions in faces. Third, the degree to which individuals distinguished between the emotion prototypes can further be predicted by the degree to which people in general distinguished between the emotions on a conceptual level as well. Further evidence can be obtained from Dissertation Project 4 – Conceptual/Perceptual (Stolier et al., 2018) on an individual level. The similarity between two personality-prototypes stemming from the same individual can be predicted by the individual's own conceptual rating about the likelihood that these two personality traits co-occur in an individual.

Together, the here presented projects highlight the technique's reliability and validity. Due to the possibility of applying the extracted prototypes to multiple faces, the technique allows the use of mixed-effects models, which further increases the generalizability of the results.

Limitations

The main limitation of the technique is that it only allows the extraction of facial information that is inherent in the used face space. The face space we used in our studies is built on predominately young White individuals. Although we have not yet tested this empirically, it is rather unlikely that the technique is able to extract facial prototypes of other ethnicities. Operating with a more diverse face space could remedy this limitation.

At this point it should also be emphasized that studies using the reverse correlation method are time consuming and rather complex. The reverse correlation method works with random stimuli and correlates participants' answers with the attributes that caused the observed choice pattern. Therefore, a variety of stimuli is needed that varies among a multitude of the dimensions the stimuli can vary on. Thus, in order to achieve reliable results, the method needs a lot of data (i.e., trials in the study). This is why I argue that our refined reverse correlation technique (or basically any reverse correlation method) is especially justified if the actual goal of the research goes beyond the mere visualization of prototypes.

Implications and Future Research

The here presented refined reverse correlation technique has many advantages that can fruitfully be used in many different domains among social psychology in particular and in psychology in general, as I will discuss in the next paragraphs.

In Moral Psychology, the character of the person who is acting plays a rather negligible role (Pizarro & Tannenbaum, 2012; Tannenbaum, Uhlmann, & Diermeier, 2011; Uhlmann, Pizarro, & Diermeier, 2015). Thus, the focus of Moral Psychology research is on the intentions, the goals, and on the outcome of the action but not on the stable actor's characteristics. However, as became apparent from the introduction of this dissertation, the stable characteristics of a person do have implications on the perception of that person. Thus, the question arises as to whether this is also true for the evaluation of a moral decision. In order to fill this gap and shed more light on the actor rather than the action of a moral

decision, Walker and Keller (2019b) extracted the prototype of someone that is perceived to be a good moral decision maker with the refined reverse correlation technique. We successfully extracted and validated the prototype of a good moral decision maker and were further able to inform that the perceived morality of an actor explains variance in the acceptance of a moral action above and beyond the intentions of the action and the outcome of the action.

Another domain in which the technique might be used beneficially is clinical psychology. Patients with borderline personality disorder tend to have a distorted self-perception as well as a distorted perception of people that are closely related to themselves but not of people who are generally known but to whom they have no personal attachments, like famous actors or successful athletes (American Psychiatric Association, 2013). There are ongoing attempts to experimentally validate existing self-report measures that can reliably distinguish between borderline personality disorder patients and the control condition (K. Schmeck, personal communication, April 4, 2019). With the use of a similar paradigm as that used in the second study by Walker and Keller (2019a), this endeavor could succeed.

The method might also be used in an adapted manner. As an example, the question of whether and how quickly facial stereotypes can be learned could be answered with a specific adaption of the technique. Let us imagine creating a random vector and taking its two opposing endpoints as the starting point for stimuli creation (i.e., starting vector). If we create, for example, 10 random vectors that we mathematically add and subtract to and from the two endpoints of the starting vector, we create two spaces that are equidistant within the individual spaces but differ systematically between each other in respect to the starting vector. The created vectors can now be applied to a set of real face photographs. In the study, participants would learn that each of the group is associated with a specific behavior (for example, cooperative versus deceptive behavior). Thus, we provide behavioral information that appears to be systematically in line with the starting random vector. In a second step,

participants would take part in a trust game. As partners, participants are presented with novel faces onto which the starting vector would be added to or subtracted from. Thus, either the facial characteristics of a novel face are in line with the group that was learned to behave cooperatively or deceptively. If the former faces are trusted more during the trust game than the latter, this would indicate that the facial characteristics that had previously been presented with cooperative behavior had been learned and used for future behavior.

A final further possibility I want to mention concerns the length of the extracted vectors. The here presented projects have in common that they all investigated the direction of the different prototype vectors within the multidimensional face space. A second piece of information that we obtained from the vectors is their length, which might be an interesting path that could be taken in future research. To illustrate, a clear selection pattern in a certain direction within the face space during the image classification task results in a longer vector compared to if the selections are more random. Thus, a longer vector might imply that a clearer internal representation has been available. The length of the average vectors can be investigated further on either the group level (e.g., are there specific personality traits that results in longer vectors than others?) or the individual level (e.g., can the length of an individual participant be related with other measures such as just world beliefs?). These and other questions could be asked and answered by including the length of the vectors in future research.

Conclusion

Faces seem to bear a vast amount of information for perceivers. We naturally and spontaneously use faces as a cue to form a first impression about someone. Furthermore, these impressions have real world implications. Thus, research in this field appears to be extremely relevant to real life. In order to investigate the specific facial characteristics that lead to these impressions, reliable and ecologically valid methods are, therefore, a necessity in the researcher's toolbox. Our data clearly suggest that the presented method is able to reliably and

validly extract a consensus about what specific stereotypes look like in a highly intuitive manner. Moreover, the here presented technique enables the extracted facial information to be applied to multiple real facial photographs and the extracted prototypes to be related with each other.

Let me refer back to the beginning where we heard about Harding and that he *looked* like a president. Today we know that Warren's administration was involved in a number of scandals and historians might say that he was one of the worst presidents the US has ever seen. Just because someone looks like a president does not guarantee that this person will *actually* be a good President. But we can provide tools that make the facial characteristics that lead to such an impression visible and therefore investigable in order to better understand the power of faces.

Literature

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, *93*, 751–763. doi: 10.1037/0022-3514.93.5.751

Agüera y Arcas, B., Todorov, A., & Mitchell, M. (2018). Do algorithms reveal sexual orientation or just expose our stereotypes? Retrieved from https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

Ahumada, A. J. (2002). Classification image weights and internal noise level estimation. *Journal of Vision*, *2*, 121–131. doi: 10.1167/2.1.8

Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. In M. D. Alicke, D. Dunning, & J. Krueger (Eds.), *The self in social judgment* (Vol. 1, pp. 85–106). New York: Psychology Press.

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology*, *20*, 1–48. doi: 10.1080/10463280802613866

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psyychiatric Pub.

Bailenson, J. N., Iyengar, S., Yee, N., & Collins, N. A. (2008). Facial similarity between voters and candidates causes influence. *Public Opinion Quarterly*, *72*, 935–961. doi: 10.1093/poq/nfn064

Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences*, *104*, 17948–17953. doi: 10.1073/pnas.0705435104

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, 19,*

187–194. doi: 10.1145/311535.311556

Brown-Iannuzzi, J. L., Dotsch, R., Cooley, E., & Payne, B. K. (2016). The relationship between mental representations of welfare recipients and attitudes toward welfare. *Psychological Science*, *28*, 92–103. doi: 10.1177/0956797616674999

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*, 305–327. doi: 10.1111/j.2044-8295.1986.tb02199.x

Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences*, *13*, 653–665. doi: 10.1016/0191-8869(92)90236-I

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychology and Personality Science*, *3*, 562–571. doi: 10.1177/1948550611430272

Dotsch, R., Wigboldus, D. H. J., Langner, O., & van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, *19*, 978–980. doi: j.1467-9280.2008.02186.x

Engell, A. D., Haxby, J. V, & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in human amygdala. *Journal of Cognitive Neuroscience*, *19*, 1508–1519. doi: 10.1162/jocn.2007.19.9.1508

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*, 878–902. doi: 10.1037/0022-3514.82.6.878

Gray, R. T. (2004). *About face: German physiognomic thought from Lavater to Auschwitz*. Wayne State University Press.

Hönekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 199–209. doi: 10.1037/0096-1523.32.2.199

Imhoff, R., Dotsch, R., Bianchi, M., Banse, R., & Wigboldus, D. H. J. (2011). Facing Europe: Visualizing spontaneous in-group projection. *Psychological Science*, *22*, 1583–1590. doi: 10.1177/0956797611419675

Imhoff, R., Woelki, J., Hanke, S., & Dotsch, R. (2013). Warmth and competence in your face! Visual encoding of stereotype content. *Frontiers in Psychology*, *4*, 1–8. doi: 10.3389/fpsyg.2013.00386

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*, 54–69. doi: 10.1037/a0028347

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*, 601–625. doi: 10.1146/annurev-psych-122414-033702

Keller, M. D., Reutner, L., Greifeneder, R., & Walker, M. (2019). *Faces evoking emotions stereotypically triggered by groups: Developing an advanced reverse correlation technique*. Manuscript under review.

Kontsevich, L. L., & Tyler, C. W. (2004). What makes Mona Lisa smile ? *Vision Research*, *44*, 1493–1498. doi: 10.1016/j.visres.2003.11.027

LeBarr, G. H. (1922). A brief analysis of president Warren G. Harding observed from the face alone. In G. H. LeBarr (Ed.), *Why you are what you are* (pp. 139–144). Boston.

Little, A. C., Burriss, R. P., Jones, B. C., & Roberts, S. C. (2007). Facial appearance affects voting decisions. *Evolution and Human Behavior*, *28*, 18–27. doi: 10.1016/j.evolhumbehav.2006.09.002

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science*, *28*, 209–226. doi: 10.1016/j.cogsci.2003.11.004

McCrae, R. R., & Costa Jr., P. T. (1997). Personality trait structure as a human universal. *American Psychologist*, *52*, 509–516. doi: 10.1037/0003-066X.52.5.509

Oliveira, M., Garcia-Marques, T., Dotsch, R., & Garcia-Marques, L. (2019). Dominance and competence face to face: Dissociations obtained with a reverse correlation approach. *European Journal of Social Psychology*. Advance online publication. doi: 10.1002/ejsp.2569

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, *18*, 566–570. doi: 10.1016/j.tics.2014.09.007

Olivola, C. Y., & Todorov, A. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, *34*, 83–110. doi: 10.1007/s10919-009-0082-1

Oosterhof, N. N., & Todorov, A. (2009). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*, 11087–11092. doi: 10.1073/pnas.0805664105

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. In *Advanced Video and Signal based Surveillance* (pp. 296–301). IEEE. doi: 10.1109/AVSS.2009.58

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), *Herzliya series on personality and social psychology. The social psychology of morality: Exploring the causes of good and evil* (pp. 91-108). Washington, DC, US: American Psychological Association. doi: 10.1037/13091-005

Porter, S., ten Brinke, L., & Gustaw, C. (2010). Dangerous decisions: The impact of first impressions of trustworthiness on the evaluation of legal evidence and defendant culpability. *Psychology, Crime and Law*, *16*, 477–491. doi: 10.1080/10683160902926141

Pound, N., Penton-Voak, I. S., & Brown, W. M. (2007). Facial symmetry is positively

associated with self-reported extraversion. *Personality and Individual Differences*, *43*, 1572–1582. doi: 10.1016/j.paid.2007.04.014

Rossion, B. (2013). The composite face illusion: A whole window into our understanding of holistic face perception. *Visual Cognition*, *21*, 139–253. doi: 10.1080/13506285.2013.772929

Rudert, S. C., Keller, M. D., Hales, A. H., Walker, M., & Greifeneder, R. (2019). *Who gets ostracized? A personality perspective on risk and protective factors of ostracism*. Manuscript in revision.

Shevlin, M., Walker, S., Davies, M. N. O., Banyard, P., & Lewis, C. A. (2003). Can you judge a book by its cover? Evidence of self-stranger agreement on personality at zero acquaintance. *Personality and Individual Differences*, *35*, 1373–1383. doi: 10.1016/S0191-8869(02)00356-2

Sigall, H., & Ostrove, N. (1975). Beautiful but dangerous: Effects of offender attractiveness and nature of the crime on juridic judgment. *Journal of Personality and Social Psychology, 31*. 410–414. doi: 10.1037/h0076472

Solomon, J. A. (2002). Noise reveals visual mechanisms of detection and discrimination. *Journal of Vision*, *2*, 105-120. doi: 10.1167/2.1.7

Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences*, *115*, 9210–9215. doi: 10.1073/pnas.1807222115

Suzuki, A., Tsukamoto, S., & Takahashi, Y. (2019). Faces tell everything in a just and biologically determined world: Lay theories behind face reading. *Social Psychological and Personality Science*, *10*, 62–72. doi: 10.1177/1948550617734616

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology*, *47*, 1249–1254. doi: 10.1016/j.jesp.2011.05.010

The United States Governement. (n.d.). Warren G. Harding. Retrieved from www.whitehouse.gov/about-the-white-house/presidents/warren-g-harding/

Todorov, A., Dotsch, R., Wigboldus, D. H. J., & Said, C. P. (2011). Data-driven methods for modeling social perception. *Social and Personality Psychology Compass*, *5*, 775–791. doi: 10.1111/j.1751-9004.2011.00389.x

Todorov, A., Loehr, V., & Oosterhof, N. N. (2010). The obligatory nature of holistic processing of faces in social judgments. *Perception*, *39*, 514–532. doi: 10.1068/p6501

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*, 1623–1626, doi: 10.1126/science.1110589.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psycholgy*, *66*, 519–545. doi: 10.1146/annurev-psych-113011-143831

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*, 455–460. doi: 10.1016/j.tics.2008.10.001

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*, 72–81. doi: 10.1177/1745691614556679

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology, 43*, 161–204. doi: 10.1080/14640749108400966

Walker, M., & Keller, M. D. (2019a). Beyond attractiveness : A multi-method approach to study enhancement in self- recognition on the Big Two personality dimensions. *Journal of Personality and Social Psychology*. Advance online publication. doi: 10.1037/pspa0000157

Walker, M., & Keller, M. D. (2019b). *Moral character matters. How act and actor (facial) characteristics impact moral judgments*. Manuscript submitted for publication.

Walker, M., Schönborn, S., Greifeneder, R., & Vetter, T. (2018). The Basel Face Database: A validated set of photographs reflecting systematic differences in Big Two and Big Five personality dimensions. *PLoS ONE*, *13*, 1–20. doi: 10.1371/journal.pone.0193190

Walker, M., & Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision*, *9*, 1–13. doi: 10.1167/9.11.12

Walker, M., & Vetter, T. (2016). Changing the personality of a face: Perceived Big Two and Big Five personality factors modeled in real photographs. *Journal of Personality and Social Psychology*, *110*, 609–624. doi: 10.1037/pspp0000064

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology, 114*, 246-257. doi: 10.1037/pspa0000098

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

Wu, X., & Zhang, X. (2016). Automated inference on criminality using face images. *ArXiv Preprint ArXiv:1611.04135*, 4038–4052.

Young, A. I., Ratner, K. G., & Fazio, R. H. (2014). Political attitudes bias the mental representation of a presidential candidate's face. *Psychological Science*, *25*, 503–510. doi: 10.1177/0956797613510717

Zebrowitz, L. A., & McDonald, S. M. (1991). The impact of litigants' babyfacedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior*, *15*, 603–623. doi: 10.1007/BF01065855

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass*, *2*, 1497–1517. doi:

10.1111/j.1751-9004.2008.00109.x

Appendices

(1) Appendix A:

Rudert, S.C., **Keller, M. D.**, Hales, A. H., Walker, M., & Greifeneder, R. (2019). *Who gets ostracized? A personality perspective on risk and protective factors of ostracism.* Manuscript in revision.

(2) Appendix B:

Keller, M. D., Reutner, L., Greifeneder, R., & Walker, M. (2019). *Faces evoking emotions stereotypically triggered by groups: Developing an advanced reverse correlation technique.* Manuscript under review.

(3) Appendix C:

Walker, M., & **Keller, M. D.** (2019). Beyond attractiveness: A multimethod approach to study enhancement in self-recognition on the Big Two personality dimensions. *Journal of Personality and Social Psychology*. Advance online publication. doi:10.1037/pspa0000157

(4) Appendix D:

Stolier, R. M., Hehman, E., **Keller, M. D.**, Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 9210-9215. doi:10.1073/pnas.1807222115

(5) Appendix E:

Curriculum Vitae

Appendix A

Rudert, S. C., **Keller, M. D.**, Hales, A. H., Walker, M., & Greifeneder, R. (2019). *Who gets ostracized? A personality perspective on risk and protective factors of ostracism.* Manuscript in revision.

Running head: PERSONALITY AND OSTRACISM

Who gets ostracized?

A personality perspective on risk and protective factors of ostracism.

Selma C. Rudert[1], Matthias D. Keller[2], Andrew H. Hales[3], Mirella Walker[2], & Rainer Greifeneder[2]

[1] University of Koblenz and Landau, Germany

[2] University of Basel, Switzerland

[3] University of Virginia, USA

Abstract

Ostracism, excluding and ignoring others, can be due to a variety of reasons. Here, we investigate the effect of personality on the likelihood of becoming a target of ostracism. Theorizing that individuals especially low in conscientiousness and/or agreeableness are at risk of getting ostracized, we tested our hypotheses within five pre-registered studies: Three experiments investigating participants' willingness to ostracize targets characterized by different personality traits, a reverse correlation face modelling study where we determined and subsequently validated the stereotypical face of an ostracized person and a survey study within a representative German data panel. In line with our hypotheses, persons low in conscientiousness and agreeableness are more likely to be intentionally ostracized by others (Studies 1 -3), represent the stereotype of an "ostracizable" person (Study 4), and report experiencing more ostracism (Study 5). Effects remained stable even after controlling for likeability of the target (Study 2 and 3). Moreover, being described as negative on one personality dimension could not be compensated by being described as positive on the other (Study 3). In exploratory analyses, we further investigated the effects of openness for experience, neuroticism and extraversion. In sum, we find evidence that personality affects the likelihood of becoming a target of ostracism, and that especially low agreeableness and conscientiousness represent risk factors.

Keywords: ostracism, personality, agreeableness, conscientiousness, person perception

Who gets ostracized?

Personality as a risk and protective factor of social ostracism.

Imagine working on a group project together with some colleagues, when you receive an email from a colleague who would be interested in joining your project. You could need some additional support and your colleague is generally a nice person. However, you know from past experience that he often shows up late to meetings, is unreliable when it comes to deadlines, and his work is often not as diligent as it should be. What would your answer to your colleague be, or would you even answer his email at all?

If you decide to decline your colleague's request to join the group project, or if you even ignore his email altogether, you have ostracized him (Williams, 2009). Ostracism is a common occurrence in everyday life, such that on average, individuals report at least one or two incidences per day where they have ignored or excluded another person as well as have been ignored or excluded by another person (Nezlek, Wesselmann, Wheeler, & Williams, 2012, 2015). Since ostracism is often a painful and threatening experience (Rudert & Greifeneder, 2016; Rudert, Hales, Greifeneder, & Williams, 2017; Williams, 2009), research investigating potential antecedents of ostracism is a highly important endeavor. Here we investigate perceived personality dispositions as a set of factors which can either put individuals at a higher risk of getting ostracized, or protect them from becoming a target. Previous research on the relation between personality and ostracism has mainly investigated the self-reported experience of the target (Wu, Wei, & Hui, 2011). Using self-reported experience generally precludes clearly identifying the cause of the identified relationships. For instance, other individuals may decide to ostracize the target in response to behavioral manifestations of the target's personality that are picked up during interactions. Alternatively, the target's personality dispositions might affect their perceptions and interpretations of what exactly constitutes an instance of ostracism. Or there

might even be reversed causal effects, such that the experience of ostracism affects the personality of the ostracized targets (Hales, Kassner, Williams, & Graziano, 2016; Nielsen, Glasø, & Einarsen, 2017). Here, we take a different approach to investigate the impact of personality dispositions by focusing on the motivations of the individuals who ostracize others (the so-called *sources* of ostracism). Using an experimental approach allows for causal interpretation of the effects of the targets' personality on the sources' intentions to ostracize.

Motivated Ostracism

To investigate potential antecedents of ostracism, it is logical to consider which factors motivate groups and individuals to ostracize others in the first place. It is often assumed that the sources of ostracism are mean and act out of malicious motives and selfishness (Rudert, Sutter, Corrodi, & Greifeneder, 2018), or simply because they do not like the ostracized target. However, in reality, that might often not be the case. Individuals need to be careful if they choose to ostracize others, because in many situations, the general norm is to include others and to let them join in activities and groups if they wish to (Rudert & Greifeneder, 2016; Wesselmann, Wirth, Pryor, Reeder, & Williams, 2013). If sources ostracize others despite this inclusion norm, they might easily end up being devalued by others or even punished for their behavior (Güroğlu, Will, & Klapwijk, 2013; Over & Uskul, 2016; Rudert, Ruf, & Greifeneder, 2018; Rudert, Sutter, et al., 2018; Will, Crone, van den Bos, & Güroğlu, 2013). Consequently, many studies have demonstrated that individuals feel uncomfortable when they ostracize others without having a plausible reason for it (Legate, DeHaan, Weinstein, & Ryan, 2013).

As a consequence, in many situations individuals will only revert to ostracism when they have a strong motive to do so and/or can assume that others will approve of their decision to ostracize others. From an evolutionary perspective, Kurzban and Leary (2001) have argued that ostracism primarily occurs if individuals are perceived as bad exchange partners. This is the case,

for instance, when (a) the ostracized target violates group norms, or (b) the ostracized target represents a burden for the sources (Kurzban & Leary, 2001; Wesselmann, Wirth, Pryor, Reeder, & Williams, 2015).

In case of (a), the target has repeatedly violated either general social norms or specific group norms, for instance by acting particularly rude, uncooperative, or ignoring specific agreements that the group made. Such norm violations threaten the harmony within a group as they create discord and increase the chance that other group members follow the negative example and start deviating from the norms as well, which would then destabilize the group and decrease cooperation (Ditrich & Sassenberg, 2016; Kerr & Levine, 2008; Scheepers, Branscombe, Spears, & Doosje, 2002). Thus, individuals may choose to ostracize targets with the goal to punish them and ultimately make them change their undesirable behavior, and thereby protect the group and its stability from being undermined by normlessness.

As for (b), some individuals may adhere to social norms, but nevertheless represent an inconvenient burden for a group. Groups often aim to achieve certain goals (McGrath, 1984), and some group members may be more useful in achieving these goals than others. And while it is a strength of groups that their members can complement each other and compensate for each other's weaknesses, a person that lacks either the skill or the motivation or both to make a meaningful contribution to a group effort can slow the group down substantially (Wesselmann, Williams, & Wirth, 2014; Wesselmann et al., 2013). Groups and their members may thus be motivated to exclude a person that they perceive to be an underperformer and burdensome, in order to keep up the group's performance (Wesselmann et al., 2014).

To sum up, an important reason for individuals to ostracize others thus is that ostracism serves as a social control mechanism, which ensures both the stability as well as the functionality of a group (Kurzban & Leary, 2001). Whether an individual is likely to violate group norms or

turns out to be a burden for the respective group might partly depend on characteristics of the group or the general situation. Some groups are more or less rigid in enforcing their norms (e.g., Gelfand et al., 2011) or have differences in proficiency levels and expectations for members. For instance, a player might easily be excluded from a professional football team if she does not meet the high standards regarding performance and/or discipline, whereas at the same time, she might be well accepted in a group that plays only occasionally for recreational purposes. However, there might also be general characteristics of individuals which makes them more or less likely to be excluded from groups, and independent of the particular social context. In what follows, we discuss how an individual's personality can affect the likelihood that this person is ostracized.

Dispositional Influences on the Likelihood of Becoming a Target of Ostracism

In our research, we focused on the so-called Big Five of personality, namely conscientiousness, agreeableness, neuroticism, openness for experience, and extraversion (Costa & McCrae, 1992). Studies that have investigated relations between the Big Five and workplace harassment or workplace ostracism have found negative correlational relationships with conscientiousness, agreeableness, and extraversion, as well as a positive correlation with neuroticism (Nielsen et al., 2017; Wu et al., 2011). In addition to being confined to a workplace setting, however, most of these studies were based on self-reports of the targets and thus do not allow for causal conclusions. Nielsen and colleagues (2017) discuss three different ways to account for the correlation between target personality and the self-reported experience of workplace harassment, which are also likely to apply to ostracism experiences in general: First, via a *target-behavior mechanism,* meaning that targets may provoke harassment via their own personality, which manifests in the behavior they show towards others (especially for highly visible traits like extraversion; Vazire, 2010). In terms of ostracism, this would be comparable to the mechanism outlined above, namely the sources excluding the targets intentionally because the

targets' behavior suggests certain personality characteristics that make the targets appear as bad exchange partners. Second, via *negative perceptions*, meaning that individuals with certain personality disposition, such as neuroticism, might be more likely to interpret negative events as harassment. In terms of ostracism, it has also been shown that the perceptions and interpretations of an ostracism episode are highly important as they can change the experience and subsequent reactions to ostracism (Downey, Mougios, Ayduk, London, & Shoda, 2004; Rudert & Greifeneder, 2016; Wirth, Lynam, & Williams, 2010; Zadro, Boland, & Richardson, 2006). Third, via a *reverse causality mechanism,* meaning that the targets' dispositions might change as a result of being exposed to harassment for a prolonged time. Again, in terms of ostracism, the existence of such vicious circles has also been shown, with ostracized individuals often acting more disagreeable, defensive, aggressive, or withdrawn (Downey, Freitas, Michaelis, & Khouri, 1998; Downey et al., 2004; Hales et al., 2016; Ren, Wesselmann, & Williams, 2016; Twenge, Baumeister, DeWall, Ciarocco, & Bartels, 2007). The present contribution mostly focuses on the target-behavior mechanism, with the aim to establish which perceived personality dispositions make the *sources* intentionally want to ostracize a target. Given this focus, we will limit our review and theorizing that are relevant for the target-behavior mechanism. In what follows, we discuss predictions separately for each of the Big Five traits.

Conscientiousness describes the tendency of individuals to act in an efficient, organized, planful, reliable, responsible, and thorough way (McCrae & John, 1992). Several studies as well as meta-analyses showed that conscientiousness is the strongest predictor of productivity and performance both in the job as well as in academic achievements (Barrick & Mount, 1991; Hurtz & Donovan, 2000; O'Connor & Paunonen, 2007; Poropat, 2009; Rothmann & Coetzer, 2003). It seems that "individuals who exhibit traits associated with a strong sense of purpose, obligation, and persistence generally perform better than those who do not" (Barrick & Mount, 1991, p. 18).

As groups often pursue certain goals, individuals who are useful in reaching these goals are likely to be valuable members of this group. In contrast, it is often said that "a chain is only as good as its weakest link," meaning that an underperforming group member can slow the group down and undermine team performance. As a consequence, it has been demonstrated that group members who are burdensome and decrease the group's performance are more likely to become targets of ostracism (Wesselmann et al., 2014; Wesselmann et al., 2013). Given the strong link between conscientiousness and performance, it thus appears likely that individuals who are low in conscientiousness are at a higher risk of becoming targets of ostracism.

Agreeableness, one of the two personality dimensions that strongly relate to social interaction, describes the tendency of individuals to act in an appreciative, kind, generous, forgiving, sympathetic, and trusting way towards others (McCrae & John, 1992). Agreeableness is linked to prosocial motivation (Graziano & Eisenberg, 1997) and thus, agreeable persons are less likely to be uncooperative or violate social norms (Berry, Ones, & Sackett, 2007; Graziano, Habashi, Sheese, & Tobin, 2007; Kagel & McGee, 2014). In contrast, this means disagreeable people are more likely to violate group norms, and destroy harmony and group cohesion, which should make them less trustworthy and more likely to become targets of ostracism. This process has been demonstrated by Hales, Kassner, Williams, and Graziano (2016) in a series of studies, showing both that self-rated agreeableness is negatively related to self-rated ostracism, but also that individuals report higher intentions to ostracize a person described as disagreeable.

Neuroticism describes the tendency to act in an anxious, self-pitying, tense, touchy, unstable, and worrying way (McCrae & John, 1992). On the one hand, it is easy to see how individuals high in neuroticism may become a burden for a group, not necessarily because of their performance, but because they may require more attention and thus be a strain for group interactions as well (Milam, Spitzmueller, & Penney, 2009). Even if group performance is not

affected, the experience of interacting with a neurotic group member may be sufficiently unpleasant to cause people to exclude and ignore this person altogether. On the other hand, individuals who are anxious and unstable may also be perceived as particularly vulnerable and thus in the need of protection of a group. Ostracizing a vulnerable person might be evaluated as particularly cruel and unfair by other group members or outsiders (Rudert, Reutner, Greifeneder, & Walker, 2017) and as a consequence, groups that ostracize a vulnerable person nevertheless might risk the anger, devaluation, or even punishment of others (Rudert, Ruf, et al., 2018; Rudert, Sutter, et al., 2018).

Openness to Experience describes the tendency to act in an artistic, curious, imaginative, insightful, and original way and have a wide range of interests (McCrae & John, 1992). Again, a person acting in such a fashion may be perceived as an attractive and interesting interaction partner and useful for group performance (especially concerning the *intellect* facet of openness), at least if creative solutions is what a group is aiming for. On the other hand, openness is also connected to pursuing unusual ideas or demonstrating unconventional behavior, which may be perceived as deviations from group norms which, again, might threaten group harmony as well as stability. Moreover, individuals high in openness are often drawn towards new situations and (social) contexts. Interestingly, irrespective of the presently discussed target behavior mechanism, this might put individuals high in openness more often in situations in which they could potentially be ostracized. Given this higher base rate of situations, they might thus be statistically also more likely to be rejected or ostracized compared to individuals who are more likely to follow more familiar patterns.

Finally, similar to agreeableness, *Extraversion* is closely related to social interactions and describes the tendency to act towards others in an active, assertive, energetic, enthusiastic, outgoing, and talkative fashion (McCrae & John, 1992). Since extraversion is often perceived as

a higher interest in social activities and interactions as well as a higher quantity of social interactions (Ashton, Lee, & Paunonen, 2002), and since extraverts tend to be more popular among peers (Jensen-Campbell et al., 2002), on first thought it seems plausible to assume a negative relation between extraversion and the risk of becoming a target of ostracism as well. However, there are several reasons why one may not expect such a direct link: First, similar to individuals high in openness, extraverts might be more likely to encounter social situations and initiate social interactions with new partners (Snyder & Gangestad, 1982), which means that the base rate of situations in which these individuals can experience ostracism is potentially higher than for individuals who stick to their close social circle. Second, even though an introverted individual might be less assertive and initiate social connections less often, that does not imply that others would deliberately decide to ostracize them or exclude them from activities they wish to join, especially as society becomes more aware of the idea of accommodating introverts (e.g., Cain, 2013). Still, highly introverted individuals might be at risk to become a target of involuntary or oblivious ostracism (Lindström & Tobler, 2018; Williams, 1997), or in other words, they might simply be overlooked on accident by others.

Taken together, from a theoretical perspective as well as from the empirical evidence, there are strong reasons to argue for a negative effect of both conscientiousness and agreeableness on the risk of becoming a target of ostracism. We thus assume that individuals who are *low in conscientiousness* are more likely to become targets of ostracism because they may be perceived as a burden for group performance, and that individuals who are *disagreeable* are more likely to become targets of ostracism, because they represent a threat to group harmony and group norms. Because theoretical predictions for openness, neuroticism and extraversion are not as clear-cut as for conscientiousness and agreeableness, we decided to look at these remaining Big Five dimensions in an exploratory fashion, as pre-registered on AsPredicted.org.

Overview of the Studies

In the present contribution, we argue for effects of perceived target personality on the likelihood to be ostracized by others that are due to a *target-behavior mechanism*, that is, the sources ostracize the targets because of their personalities. Consequentially, there should be a direct causal link from sources' perception of the target's personality on their likelihood to ostracize the target. Studies 1 - 3 test for a causal effect of perceived target conscientiousness and agreeableness on ostracism intentions, while controlling for liking (Studies 2 and 3). We hypothesized that low conscientiousness should increase the likelihood of becoming a target of ostracism. Moreover, replicating the findings of Hales and colleagues (2016), we further assumed that low agreeableness would increase the likelihood of becoming a target of ostracism. In an exploratory fashion, in Studies 1 and 2 we also investigate potential causal effects of the other three personality dimensions (extraversion, openness, neuroticism) on ostracism intentions without having a priori hypotheses for the exact nature of this effect. Study 3 specifically tests for interactions between conscientiousness and agreeableness.

Study 4 investigates the effect of personality on ostracism intentions with a different, more subtle method, namely via face perception. Research has demonstrated that there is a strong social consensus how the face of a person with a certain personality appears (Walker, Schönborn, Greifeneder, & Vetter, 2018; Walker & Vetter, 2016) and that individuals also intuitively base their (moral) judgments upon these facial cues (Funk, Walker, & Todorov, 2017; Rudert, Reutner, et al., 2017). Against this background, we expect a social consensus of how a person that is likely to be ostracized stereotypically appears, and that this consensus will bear similarity to the consensus for facial appearance of a person with certain personality characteristics (e.g., a careless and disagreeable person). We investigate this hypothesis in Study 4, using a reverse correlation paradigm.

Finally, if there is a substantial effect of target personality on ostracism, one could expect to find an association between self-reported personality and self-reported feelings of ostracism that can be identified in the real world outside of the laboratory. There are some studies regarding ostracism and harassment specifically in the workplace that hint to such a relation (Nielsen et al., 2017; Wu et al., 2011), however, it appears important to go beyond this preliminary evidence and demonstrate that the hypothesized relations can be demonstrated a) independent of a specific context such as the workplace and b) in a nation-wide representative sample. In Study 5, we thus investigate the relationship of the Big Five Personality Dimensions with subjectively experienced ostracism in a representative, longitudinal panel (the innovation sample of the German socio-economic panel; SOEP-IS). Specifically, we test whether (prospectively measured) conscientiousness and agreeableness negatively predict self-reported ostracism.

Study 1

In Study 1, we manipulated personality of a presented target and measured the potential sources' intention to ostracize the target. We predicted that individuals report higher ostracism intentions for targets who were described as either low in conscientiousness or agreeableness. It should be noted that ostracism that is perceived as malicious and unfair can easily result in devaluation and punishment by others (Rudert, Ruf, et al., 2018; Rudert, Sutter, et al., 2018), and also that people are aware of, and sensitive to, the pain of others who are ostracized (Coyne, Nelson, Robinson, & Gundersen, 2011; Wesselmann, Bagg, & Williams, 2009). Thus, ostracism will usually not be used light-heartedly. Specifically, individuals will likely ostracize others when they feel they have a valid reason to do so, namely if they have to protect themselves and their group from bad exchange partners (Kurzban & Leary, 2001). Consequentially, we predicted that rather than following a strictly linear function, intentions to ostracize would increase specifically

as a function of low conscientiousness as well as low agreeableness, compared to both high conscientiousness/agreeableness as well as a neutral control condition.

Method

Participants and design. Participants were recruited online from Prolific Academic (US Americans only) for a payment of £0.40. Based on the studies from Hales and colleagues (2016), we had initially calculated the sample size such as to detect a large-sized main effect of each personality dimension on participants' ostracism intentions (f = .40, power = .80, required n = 304). However, data analysis from a pretest showed that while we detected an effect of agreeableness with a comparable effect size as Hales and colleagues, this effect was much larger than the effect of any other personality dimension, so power was likely too low to detect effects on the other dimensions that appeared to be more of a medium size. Moreover, some of the manipulation checks for our initial personality descriptions of the target person were not satisfactory. We thus slightly re-phrased some of the descriptions, which were then pre-tested separately and this time performed adequately. Additionally, based on these initial findings we re-calculated the sample size, this time to detect medium-sized main effects (f = .25, power = .80, required n = 579). Adding a buffer of 25 percent, we ran another, adequately powered study with a different sample of 801 participants on Prolific Academic, excluding all participants who indicated that their data should not be used (79) as well as participants who already participated in the pretest (7). The final sample thus consisted of 715 participants (335 females, 3 no specified gender; M_{age} = 34.67, SD = 24.90). Here, we report only the data of the second, adequately powered sample. Participants were randomly assigned to eleven conditions (conscientiousness high vs. low; agreeableness high vs. low; neuroticism high vs. low; openness high vs. low; extraversion high vs. low; and a control condition). The study was also preregistered on AsPredicted.org, see https://aspredicted.org/blind.php?x=sb36k8.

Materials and procedure. The study's procedure was adapted from Hales and colleagues (2016). Participants read a vignette that described a student named Mason. The basic version of each vignette contained no information about Mason's personality and was the same for all groups.

Mason is a 19 year old Sophomore student. He works as a part time job at a nearby restaurant. In his free time, he likes to watch movies, listen to music, and go outdoors. In a typical day, Mason goes to classes and afterwards spends some time on his computer. After dinner, he usually watches TV shows. His favorites are crime series, but he also enjoys quiz shows.

To make sure there would be no floor or ceiling effects of personality ratings of the basic version, we pretested the vignette with 15 participants on Prolific Academic (9 female; M_{age} = 29.60, *SD* = 7.50) who rated Mason on the five personality dimensions (e.g., *"Mason is disagreeable – agreeable"*, 7-point scale). Mason was rated as fairly average on the five dimensions (conscientiousness: *M* = 4.93, *SD* = 1.39; agreeableness: *M* = 4.60, *SD* = .99; neuroticism *M* = 3.20, *SD* = 1.42; openness: M = 4.67, *SD* = .98; extraversion: *M* = 3.53, *SD* = 1.25). In the actual study, participants in the control condition received the basic version of the vignette. In the remaining 10 groups, Mason was additionally described as either being high or low in one of the manipulated Big Five personality dimensions within the vignette, see *Table 1*.

Table 1
Manipulation of the Personality Dimensions in Study 1 and 2.

Personality Dimension	*Low*	*High*
Conscientiousness	Mason tends to be a lazy, chaotic person and an unreliable and careless worker.	Mason tends to be a diligent, well-organized person and a reliable and precise worker.
Agreeableness	Mason tends to be a cold, untrusting, and uncaring person.	Mason tends to be a warm, trusting, and caring person.
Neuroticism	Mason tends to be a relaxed, confident, and cheerful person.	Mason tends to be an anxious, insecure, and moody person.
Openness	Mason tends to be an unimaginative person who likes to think in familiar patterns and prefers to do things in a routine way.	Mason tends to be an ingenious person who likes to come up with new ideas and prefers to do things in an inventive way.
Extraversion	Mason tends to be a quiet, calm, and reserved person.	Mason tends to be a talkative, energetic, and outgoing person.

After reading the vignette, participants first answered a manipulation check about the respective personality dimension that was manipulated in their condition (e.g., *"Mason is disagreeable – agreeable"*, 7-point scale) as well as a control question (*"Mason likes crime series – hates crime series",* 7-point scale). In the control condition, participants rated Mason on all five personality dimensions. Participants were then asked to imagine that Mason wanted to join a club they already belonged to and then reported their intention to ostracize Mason on a scale consisting of seven items (Cronbach's $\alpha = .89$; exemplary item: "I might find myself ignoring Mason", 1 = completely disagree, 5 = completely agree; Hales et al., 2016). After providing demographic information, participants were thanked and paid.

Results

Manipulation checks. We tested whether participants perceive the manipulation of the personality dimension using five one-way ANOVAs with the manipulated personality dimension (high vs. control vs. low) on the respective manipulation check. All ANOVAs revealed significant differences with regard to the manipulated trait (Conscientiousness: $F(2,197) = 286.51$, $p < .001$, $\eta^2 = .74$; Agreeableness: $F(2,198) = 134.07$, $p < .001$, $\eta^2 = .58$; Neuroticism:

$F(2, 194) = 142.00$, $p < .001$, $\eta^2 = .59$; Openness: $F(1, 197) = 51.59$, $p < .001$, $\eta^2 = .34$; Extraversion: $F(2, 186) = 60.85$, $p < .001$, $\eta^2 = .40$). Bonferroni-corrected post-hoc analyses showed that for all five dimensions, the high condition was significantly greater than the control condition, and the low condition was significantly smaller than the control condition, smallest $p = .003$. See *Table 2* for the descriptive statistics.

Table 2
Manipulation Checks for Study 1

Personality Dimension	*Low*	*Control*	*High*
Conscientiousness	1.65 (1.36) [a]	5.29 (1.34) [b]	6.50 (0.86) [c]
Agreeableness	2.85 (1.92) [a]	5.62 (1.17) [b]	6.58 (0.70) [c]
Neuroticism	1.69 (0.98) [a]	2.49 (1.24) [b]	5.22 (1.47) [c]
Openness	3.42 (1.76) [a]	5.13 (1.39) [b]	5.96 (1.18) [c]
Extraversion	2.28 (1.45) [a]	3.63 (1.66) [b]	5.57 (1.82) [c]

Note. Means (and standard deviations) on the manipulation checks as a function of the manipulated personality dimension. The letters a - b represent significant differences between groups; all values in the same row that share the same letter do not differ significantly from each other, values with different letters do.

Dependent variables. Five one-way ANOVAs (manipulated personality dimension: high vs. control vs. low) on intentions to ostracize Mason revealed significant differences for Conscientiousness, $F(2,197) = 38.18$, $p < .001$, $\eta^2 = .28$; Agreeableness, $F(2,198) = 37.38$, $p < .001$, $\eta^2 = .27$; Neuroticism, $F(2, 194) = 11.45$, $p < .001$, $\eta^2 = .11$; and Openness, $F(2, 197) = 5.31$, $p = .006$, $\eta^2 = .05$. Manipulating Extraversion did not predict intentions to ostracize, $F(2,186) = 0.01$, $p = .986$, $\eta^2 < .01$. Bonferroni-corrected post-hoc tests revealed that the differences mostly derived from the negative pole differing significantly from both the positive pole and the control group, largest $p = .016$. Participants reported a higher intention to ostracize persons who were careless, disagreeable, emotionally unstable, and close-minded, see *Table 3* and *Figure 1* for the descriptive statistics. In contrast, there were no significant differences between the positive pole and the control group, smallest $p = .659$.

Table 3

Results for Study 1

Personality Dimension	*Low*	*Control*	*High*
Conscientiousness	2.58 (.87) [a]	1.67 (.73) [b]	1.54 (.60) [b]
Agreeableness	2.65 (.92) [a]	1.67 (.73) [b]	1.61 (.68) [b]
Neuroticism	1.52 (.64) [a]	1.67 (.73) [a]	2.10 (.78) [b]
Openness	2.03 (.73) [a]	1.67 (.73) [b]	1.68 (.67) [b]
Extraversion	1.68 (.60) [a]	1.67 (.73) [a]	1.69 (.75) [a]

Note. Means (and standard deviations) of participant's intention to ostracize Mason as a function of the manipulated personality dimension. Note that the control group was the same for all personality dimensions (no information about personality provided). The letters a - b represent significant differences between groups; all values in the same row that share the same letter do not differ significantly from each other, values with different letters do.

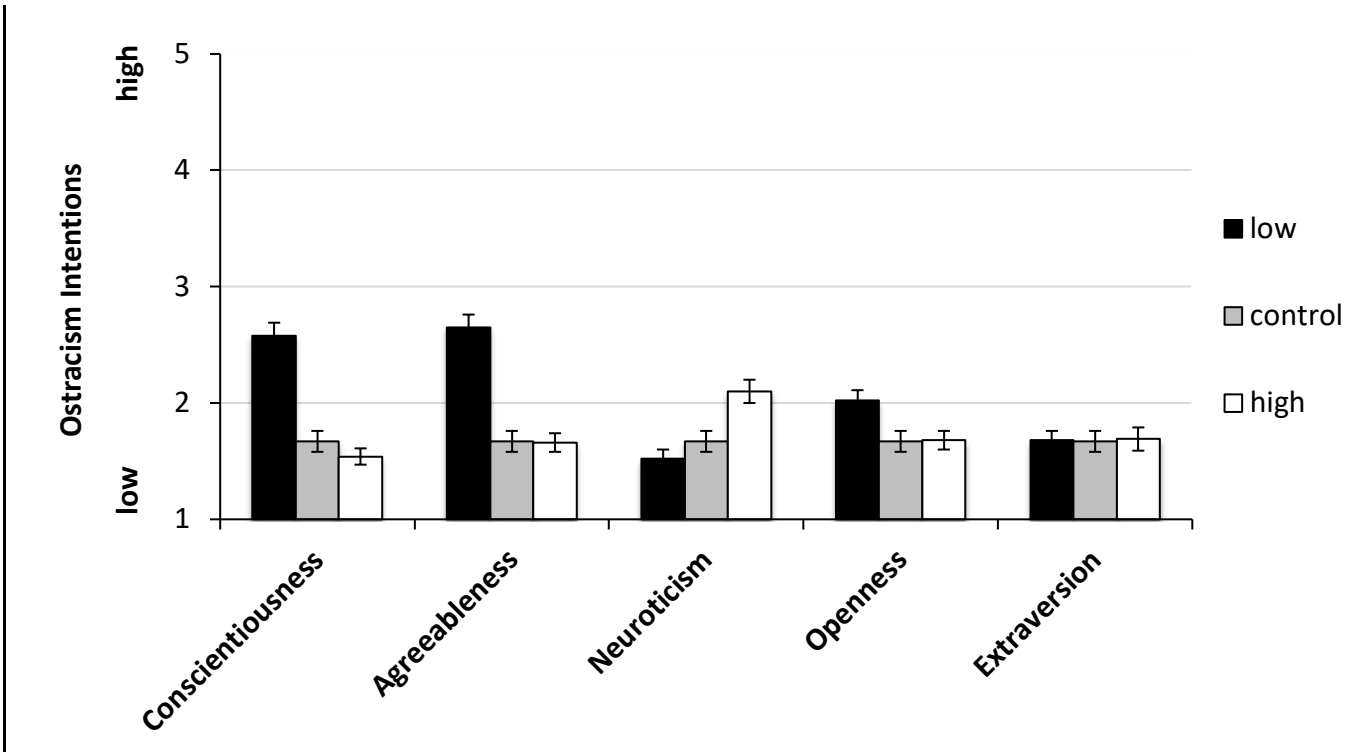


Figure 1. Mean ostracism intentions (with standard errors) as a function of the manipulated personality dimension in Study 2. Note: There was only one control condition, it is displayed multiple times above for ease of interpretation.

Discussion

In line with our hypotheses, we show that a target who is low in conscientiousness or low in agreeableness elicited greater intentions to ostracize the target. This effect was specific to the negative disposition; there was no significant difference between the neutral control condition and a positive description on the respective dimension. As an exploratory analysis, we also investigated the other three Big Five dimensions. A description of the target as high in neuroticism or low in openness also increased participant's intentions to ostracize, even though the effect was much smaller than the effect of low conscientiousness and disagreeableness. It should be noted that we cannot rule out the possibility that manipulating one personality dimension also changed participant's perception of other dimensions. In fact, spillover effects are likely given that the Big Five are naturally correlated with each other (Soto & John, 2017) and thus it makes sense that participants would infer from one personality dimension to others. We further address this issue in Study 3.

While participants were more inclined to ostracize persons who were careless, disagreeable, emotionally unstable, and close-minded, it appears that being explicitly conscientious, agreeable, emotionally stable, and open-minded did not offer any additional protection. This does not appear to be due to a failure to detect differences between control condition and high levels of these traits; the manipulation checks uniformly showed that people perceived the personality differences. Given that the reported intention to ostracize others was low on average, the finding could potentially be due to a floor effect. However, the resulting pattern also aligns with our theoretical assumptions that individuals will not decide to ostracize others easily, and will thus primarily ostracize others when they feel they have a valid reason to do so, such as the target being a particularly bad exchange partner.

Study 2

As a next step, we investigated more closely what drives the effect of personality on ostracism intentions. One plausible mediator is *liking.* Previous research has demonstrated medium to strong positive relations between how much a person is liked and how that person is rated on conscientiousness, agreeableness, openness, and extraversion, as well as a negative relation between liking and neuroticism (Leising, Erbs, & Fritz, 2010). In addition, Hales and colleagues (2016) also showed that disliking predicts ostracism intentions and found it to be a strong mediator of the relation between ostracism and agreeableness.

While liking is thus a plausible mediator for the effect of personality on ostracism, there are also reasons to assume that the relationship between personality and ostracism does not exclusively depend on liking alone (Hales et al., 2016). Study 2 thus had both the aim to replicate the findings from Study 1 as well as to test whether the effect of personality dispositions on ostracism is fully mediated by liking, or whether there is unique variance not explained by liking of the ostracized target.

Method

Participants and design. Participants were recruited online from Prolific Academic (US Americans only) for a payment of £0.40. As in Study 1, we recruited 809 participants on Prolific Academic, excluding all participants who indicated that their data should not be used (9). The final sample thus consisted of 800 participants (345 females, 3 no specified gender; M_{age} = 32.35, *SD* = 11.38). Participants were randomly assigned to eleven conditions (Conscientiousness high vs. low; Agreeableness high vs. low; Neuroticism high vs. low; Openness high vs. low; Extraversion high vs. low; and a control condition). The study was also preregistered on AsPredicted.org, see http://aspredicted.org/blind.php?x=qi7vv5

Materials and procedure. The study's procedure was similar to Study 1. After answering the questions about intentions to ostracize (Cronbach's α = .89), participants answered how much they liked Mason on a five-item scale (Cronbach's α = .87; exemplary item: "Mason is likeable", 1 = completely disagree, 5 = completely agree; Hales et al., 2016).

Results

Manipulation checks. We tested participants' perception of the manipulated personality dimensions (high vs. control vs. low) using five one-way ANOVAs on the respective manipulation check. All ANOVAs revealed significant differences with regard to the manipulation (Conscientiousness: $F(2,217) = 736.08$, $p < .001$, $\eta^2 = .87$; Agreeableness: $F(2,215) = 146.43$, $p < .001$, $\eta^2 = .58$; Neuroticism: $F(2, 214) = 88.14$, $p < .001$, $\eta^2 = .45$; Openness: $F(1, 216) = 108.45$, $p < .001$, $\eta^2 = .50$; Extraversion: $F(2, 214) = 138.33$, $p < .001$, $\eta^2 = .56$). Bonferroni-corrected post-hoc analyses testing the differences between a high and a low value on the respective dimension compared to the control group were all significant as well, all $p < .001$. See *Table 4* for the descriptive statistics.

Table 4
Manipulation Checks for Study 2

Personality Dimension	*Low*	*Control*	*High*
Conscientiousness	1.33 (.55) [a]	5.11 (1.13) [b]	6.50 (.73) [c]
Agreeableness	2.88 (1.68) [a]	5.13 (1.14) [b]	6.42 (.84) [c]
Neuroticism	1.96 (1.14) [a]	2.89 (1.43) [b]	5.08 (1.73) [c]
Openness	3.19 (1.57) [a]	4.69 (.93) [b]	6.15 (1.06) [c]
Extraversion	2.19 (1.04) [a]	3.33 (1.35) [b]	5.90 (1.66) [c]

Note. Means (and standard deviations) on the manipulation checks as a function of the manipulated personality dimension. The letters a - b represent significant differences between groups; all values in the same row that share the same letter do not differ significantly from each other, values with different letters do.

Dependent variables. *Ostracism Intentions.* Five one-way ANOVAs (manipulated personality dimension: high vs. control vs. low) on intentions to ostracize Mason revealed

significant differences for Conscientiousness, $F(2,214) = 50.09$, $p < .001$, $\eta^2 = .32$; Agreeableness, $F(2,215) = 69.18$, $p < .001$, $\eta^2 = .39$; Neuroticism, $F(2, 214) = 5.50$, $p = .005$, $\eta^2 = .05$; and Openness, $F(2, 216) = 4.77$, $p = .009$, $\eta^2 = .04$. Manipulating Extraversion did not affect intentions to ostracize, $F(2,214) = 0.43$, $p = .650$, $\eta^2 < .01$. Again, Bonferroni-corrected post-hoc tests revealed that the differences derived from the negative pole differed significantly from both the positive pole and the control group, largest $p = .037$. Participants reported a higher intention to ostracize persons who were careless, disagreeable, emotionally unstable, and close-minded. In contrast, there were no significant differences between the positive pole and the control group, smallest $p = .891$, so that being explicitly agreeable, conscientious, emotionally stable, and open-minded did not offer any additional protection. See *Table 5* for the descriptive statistics.

Table 5
Results for Study 2

Personality Dimension	*Dependent Variable*	*Low*	*Control*	*High*
Conscientiousness	Ostracism Intentions	2.60 (.83) [a]	1.59 (.65) [b]	1.52 (.69) [b]
	Liking	2.77 (.71) [a]	3.62 (.59) [b]	3.88 (.59) [c]
Agreeableness	Ostracism Intentions	2.72 (.94) [a]	1.59 (.65) [b]	1.47 (.44) [b]
	Liking	2.67 (.84) [a]	3.62 (.59) [b]	3.87 (.62) [b]
Neuroticism	Ostracism Intentions	1.52 (.63) [a]	1.59 (.65) [a]	1.88 (.75) [b]
	Liking	3.81 (.61) [a]	3.62 (.59) [a]	3.27 (.77) [b]
Openness	Ostracism Intentions	1.87 (.70) [a]	1.59 (.65) [b]	1.57 (.63) [b]
	Liking	3.39 (.68) [a]	3.62 (.59) [a]	3.95 (.57) [b]
Extraversion	Ostracism Intentions	1.66 (.61) [a]	1.59 (.65) [a]	1.68 (.64) [a]
	Liking	3.79 (.55) [a]	3.62 (.59) [a]	3.75 (.62) [a]

Note. Means (and standard deviations) of participant's intention to ostracize Mason as a function of the manipulated personality dimension. Note that the control group was the same for all personality dimensions (no information about personality provided). The letters a - b represent significant differences between groups; all values in the same row that share the same letter do not differ significantly from each other, values with different letters do.

Liking. Conscientiousness, agreeableness, neuroticism, and openness also significantly affected how much participants liked Mason (Conscientiousness, $F(2,214) = 60.77$, $p < .001$, $\eta^2 =$

.36; Agreeableness, $F(2,215) = 61.25$, $p < .001$, $\eta^2 = .36$; Neuroticism, $F(2, 214) = 12.52$, $p < .001$, $\eta^2 = .11$; and Openness, $F(2, 216) = 15.48$, $p < .001$, $\eta^2 = .13$). Extraversion did not affect liking, $F(2,214) = 1.69$, $p = .187$, $\eta^2 = .02$.

Looking at the pattern of means, liking was distributed more linearly than ostracism, intentions, see *Table 5* for the descriptive statistics. Bonferroni-corrected post-hoc tests for conscientiousness showed that participants liked Mason better in the control group compared to when he was described as careless, $p < .001$ and even better when he was explicitly conscientious compared to the control group, $p = .046$. For neuroticism and agreeableness, Mason was also liked less when he was described as emotionally unstable or disagreeable, largest $p = .006$, but not significantly more when he was described as emotionally stable or agreeable, smallest $p = .079$. For openness, Mason was liked more when describes as open compared to the control group, $p = .003$, but not less when he was described as close-minded compared to control group, $p = .085$.

ANCOVA and Mediation. When controlling for liking, intentions to ostracize Mason remained significant for conscientiousness, $F(2,213) = 16.08$, $p < .001$, $\eta^2 = .13$ and agreeableness, $F(2,214) = 24.27$, $p < .001$, $\eta^2 = .19$; but not for neuroticism, $F(2,213) = 1.60$, $p = .205$, $\eta^2 = .02$, and openness, $F(2,215) = 2.65$, $p = .073$, $\eta^2 = .02$. We additionally ran mediation analyses with PROCESS (Hayes, 2013), using 5,000 bootstrap estimates. Liking mediated the effect of conscientiousness on ostracism intentions, $b_{indirect} = -.23$, 95% CI = [-.36; -.12], though the direct effect remained significant as well, $b_{direct} = -.32$, $p < .001$, 95% CI = [-.45; -.17]. Similarly, liking mediated the effect of agreeableness on ostracism intentions, $b_{indirect} = -.23$, 95% CI = [-.38; -.11], but again, the direct effect remained significant, $b_{direct} = -.39$, $p < .001$, 95% CI = [-.53; -.25]. When running mediation analyses with neuroticism or openness as the predictor, only the indirect effects were significant, neuroticism: $b_{direct} = .09$, $p = .125$, 95% CI = [-.02; .20],

$b_{indirect}$ = .09, 95% CI = [.03; .16]; openness: b_{direct} = -.10, p = .084, 95% CI = [-.21; .01], $b_{indirect}$ = -.05, 95% CI = [-.11; .00].

Discussion

Study 2 fully replicated the results of Study 1: When Mason was described as either low in conscientiousness or low in agreeableness, participants reported stronger intentions to ostracize him. As in Study 1, high neuroticism and low openness also increased ostracism intentions while extraversion did not. Again, all detected differences were due to a *negative* description of the target compared to both the positive as well as the neutral control condition, and again this is not attributable to failing to detect the positive personality conditions as differing from the control condition. Interestingly, this pattern was different for liking: Liking increased as a linear function of the respective personality dimensions, such that individuals were liked the more conscientious, open, (and by trend, the more agreeable) they were described.

This difference between ostracism intentions and liking was also reflected in the finding that even after including liking as a control variable, the effects of agreeableness and conscientiousness on ostracism intentions remained significant. Mediation analyses, as expected, showed liking to be an important mediator of the relation between personality factors and ostracism intentions. However, particularly for conscientiousness and agreeableness, significant variance could not be explained by how much participants liked Mason. For agreeableness, meaningful parts of this variance are likely due to the tendency to *distrust* low agreeable targets as interaction partners – a factor identified to mediate this effect, controlling for liking, in earlier research (Hales et al., 2016). Thus, it seems that while liking can explain a large portion of the conscientiousness/agreeableness – ostracism link, it cannot account for the relation on its own. This finding makes sense if one considers that there might be cases in which ostracism occurs independent of liking: For instance, a careless target might be well liked in principle, but still be

excluded from a highly performance-oriented group. On the other hand, a person that is strongly disliked by others might not be ostracized because the group depends on him/her or cannot afford to exclude one of its members.

Study 3

Studies 1 and 2 investigated the effects of the Big Five on ostracism intentions independently from each other, which theoretically allows a focused test of each dimension. Yet is also conceivable that participants go beyond the information given, assuming that a person who is being described as negative on one personality dimension is also negative on others (a negative halo effect, Nisbett & Wilson, 1977). Alternatively, one could assume that a personality disposition that is perceived as positive can compensate for another negative personality dimension (Kervyn, Yzerbyt, Demoulin, & Judd, 2008), such that a person is less likely ostracized if there is at least some good in her. Consequentially, the effects may be purely additive – the more positive characteristics individuals have, the less likely they are, linearly, to be ostracized and vice versa. To test these competing predictions, in Study 3, we investigated the interaction of perceived conscientiousness and agreeableness on ostracism intentions. Since we assume that both conscientiousness and agreeableness are highly central for a person to be perceived as a good exchange partner (e.g., Fiske, Cuddy, & Glick, 2007), we predicted that ostracism intentions would increase as soon as the target would be described as *either* negative regarding conscientiousness *or* negative regarding agreeableness, regardless of how the target is described on the respective other dimension. In other words, we expected that being negative on one personality dimension cannot be compensated by being positive on the other.

Method

Participants and design. Participants were recruited online from Prolific Academic (US Americans only) for a payment of £0.40. The previous studies showed that we could expect large

effect sizes for the main effect of both conscientiousness and agreeableness on ostracism intentions (f = .40, power = .80, required N = 52). Following recent recommendations that to calculate power for an interaction hypothesis with a "knockout pattern," one should calculate four times the sample size of the original main effect (Giner-Sorolla, 2018; Simonsohn, 2014), we concluded that we would need at least data of N = 208. Adding 10% for reasons of unexpected dropouts, we thus collected a sample of 232 participants on Prolific Academic, excluding all participants who indicated that their data should not be used (1 person). The final sample thus consisted of 231 participants (108 females; M_{age} = 34.43, SD = 11.83). Participants were randomly assigned to a 2 (conscientiousness high vs. low) x 2 (agreeableness high vs. low) between-subject design. The study was preregistered on AsPredicted.org, see http://aspredicted.org/blind.php?x=vq3e7y

Materials and procedure. The study's procedure was similar to Studies 2 and 3. Participants read the vignette about Mason, who was either described as high vs. low in conscientiousness and high vs. low in agreeableness. The order of each personality description was randomly counterbalanced, such that some participants read about agreeableness first, and others read about conscientiousness first. Participants then answered the manipulation check about conscientiousness and agreeableness as well as the control/filler question about how much Mason liked crime series, reported their intention to ostracize Mason (Cronbach's α = .92) as well as how much they liked Mason (Cronbach's α = .94). After providing demographic information, participants were thanked and paid.

Results

Manipulation checks. Participants in the high conscientiousness condition perceived Mason to be more conscientious than participants in the low conscientiousness condition, $F(1,227) = 658.12$, $p < .001$, $\eta^2 = .74$ (M = 5.94, SD = 1.59 vs. M = 1.54, SD = 1.11). Moreover,

participants in the high agreeableness condition perceived Mason to be more agreeable than participants in the low agreeableness condition, $F(1,227) = 378.45$, $p < .001$, $\eta^2 = .63$ ($M = 5.94$, $SD = 1.25$ vs. $M = 2.56$, $SD = 1.50$). There was also a significant, though substantially smaller, effect of each personality dimension on the other manipulation check, such that participants assumed that Mason was more agreeable when he was described as conscientious (vs. careless), $F(1,227) = 22.52$, $p < .001$, $\eta^2 = .09$ ($M = 4.10$, $SD = 2.58$ vs. $M = 2.56$, $SD = 1.50$) and more conscientious when he was described as agreeable (vs. disagreeable), $F(1,227) = 22.42$, $p < .001$, $\eta^2 = .09$ ($M = 4.67$, $SD = 2.11$ vs. $M = 3.85$, $SD = 2.19$). There were no significant interactions, smallest $p = .135$.

Dependent variables. A two-way ANOVA (conscientiousness: high vs. low and agreeableness: high vs. low) on intentions to ostracize Mason revealed significant main effects of Conscientiousness, $F(1,227) = 48.26$, $p < .001$, $\eta^2 = .18$ and Agreeableness, $F(1,227) = 173.33$, $p < .001$, $\eta^2 = .43$, replicating the finding that both low conscientiousness and low agreeableness increased ostracism intentions. Most important however, the hypothesized interaction was significant, $F(1,227) = 12.64$, $p < .001$, $\eta^2 = .05$, see *Figure 2*. Simple main effect analyses showed that intentions to ostracize Mason were highest when he was described negative on both personality dimensions, but that being positive on either one of the two personality dimensions could not be compensate for being negative on the other: When Mason was described as conscientious, low agreeableness resulted in a substantial increase in ostracism intentions compared to high agreeableness, $F(1, 227) = 138.00$, $p < .001$, $\eta^2 = .38$ ($M = 2.91$, $SD = .77$ vs. $M = 1.34$, $SD = .38$); also, when he was described as agreeable, low conscientiousness resulted in a substantial increase in ostracism intention ($M = 2.33$, $SD = .85$) compared to high conscientiousness, $F(1, 227) = 55.38$, $p < .001$, $\eta^2 = .20$. When Mason was already described as low in conscientiousness, low agreeableness further increased ostracism intentions ($M = 3.23$, SD

= .76) but to a much smaller degree compared to when he was described as conscientious, $F(1, 227) = 46.79$, $p < .001$, $\eta^2 = .17$. On the same note, when Mason was already described as disagreeable, low conscientiousness increased ostracism intentions to a much smaller degree than when he was described as agreeable, $F(1, 227) = 5.73$, $p = .018$, $\eta^2 = .03$.

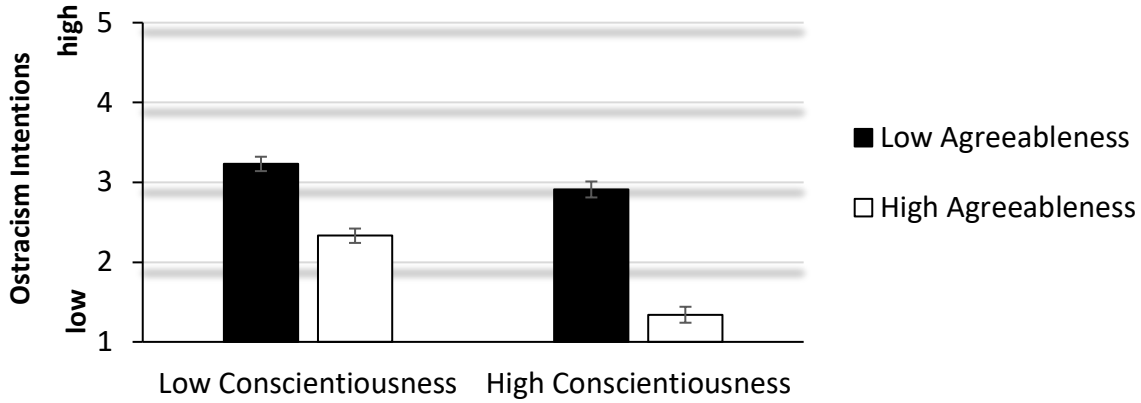


Figure 2. Mean ostracism intentions (with standard errors) as a function of manipulated agreeableness and conscientiousness in Study 3.

This pattern became more pronounced when controlling for liking. Although all main effects and interactions remained significant (conscientiousness: $F(1, 226) = 8.52$, $p = .004$, $\eta^2 = .04$, agreeableness: $F(1, 226) = 5.24$, $p = .023$, $\eta^2 = .02$, agreeableness x conscientiousness: $F(1, 226) = 4.91$, $p < .001$, $\eta^2 = .02$), the simple main effects showed that when Mason was already described negative on one personality dimension, adding a second negative personality characteristic did not significantly increase ostracism intentions, both $F < 1$. In contrast, describing an agreeable person as careless, or a conscientious person as disagreeable significantly increased ostracism intention compared to being described as positive on both dimensions, $F(1, 226) = 12.21$, $p = .001$, $\eta^2 = .05$ and $F(1, 226) = 8.42$, $p = .004$, $\eta^2 = .04$.

There was a significant main effect of the order in which the personality dimensions were presented, $F(1, 223) = 5.24$, $p = .009$, $\eta^2 = .03$, namely that participants reported higher ostracism intentions when the information about agreeableness was presented first compared to

second (M = 2.59, SD = 1.03 vs. M = 2.32, SD = .97). However, all interactions of order with the manipulated personality dimensions were not significant, smallest p = .197.

Discussion

Study 3 tested the interaction effect of conscientiousness and agreeableness on ostracism intentions. As predicted, being described as either careless or disagreeable significantly increased ostracism intentions, even if the target was described as positive on the respective other dimension. In contrast, being described as negative on both dimensions, compared to one, increased ostracism intentions only slightly, a difference that was not even significant when controlling for liking. This finding is in line with the strong significance of negative information in impression formation that has repeatedly been demonstrated in the literature on the negativity bias (Fiske, 1980; Skowronski & Carlston, 1989). From the perspective that ostracism ultimately serves the goal to eliminate bad exchange partners, it appears logical that being highly negative on one dimension cannot be compensated with being positive on the other: Ultimately, an individual that is agreeable but careless and underperforming might be just as problematic and troublesome for a group as an individual that performs well and reliably, but constantly disrupts the harmony and cohesion of the group. This focus on both the intent as well as the capability of a person is also highlighted in related models, such as the Stereotype Content Model (Fiske et al., 2007).

Another interpretation of our findings could be that the results are in fact due to some kind of a negative halo effect, as the manipulation checks show that a person that is described as disagreeable is also perceived as less conscientious and vice versa. Thus, the fact that there is no additive effect of personality dimensions on ostracism intentions could be due to participants negatively adjusting their perception of Mason's entire personality when learning that he has one negative characteristic. Given that agreeableness and conscientiousness are correlated with each

other (Soto & John, 2017), it is not surprising that descriptions of one trait colored impressions of the other. However, because the main effects on the targeted traits were much larger than the spillover effects, it is parsimonious to conclude that the observed pattern is due to the unique combination of each independent trait.

Taken together, Studies 1-3 show that individuals report higher ostracism intentions towards targets with specific personality dispositions. Particularly, and as hypothesized, low conscientiousness and agreeableness strongly affected ostracism intentions. The effects were partly mediated by liking, however, the effect of both conscientiousness and agreeableness remained significant even after controlling for liking. Exploratory analyses of the results in Studies 1 and 2 showed that high neuroticism and low openness were also associated with stronger ostracism intentions, and that this relation could be partially explained by differences in liking.

Study 4

In Studies 1 – 3, personality was manipulated directly via description of the target person and ostracism intentions were assessed rather explicitly, namely by asking participants how likely it is that they would ostracize a specific person. Since explicit preferences are sometimes prone to biases (e.g., Paulhus & Vazire, 2007), it appeared desirable to further demonstrate the effect of personality on ostracism intentions with a more subtle measure. To this end, we built on evidence demonstrating a strong cross-cultural social consensus in personality judgments from faces (Walker, Jiang, Vetter, & Sczesny, 2011; Walker & Vetter, 2016). This research finds that people generally agree that certain faces appear to convey greater or lower levels of the five personality dimensions. For example, others tend to agree that a particular person's face appears more or less conscientious. Although these judgments are not necessarily externally valid (Olivola & Todorov, 2010), this consensus allows for a different test of our hypotheses: Given a consensus about how,

for instance, a conscientious or agreeable person looks, and given our findings that conscientiousness and agreeableness are reliably associated with the likelihood of being ostracized, one might expect that a person with facial features signaling low agreeableness or conscientiousness is also perceived as a person that is more likely to be ostracized. To test this assumption, we proceed in two steps: Study 4a uses an image classification task (e.g., Mangini & Biederman, 2004; Dotsch et al., 2008) and a statistical face modeling technique (Walker & Vetter, 2016) to find out and visualize how people mentally represent the face of a person likely to be ostracized. Study 4b then presents this face to another sample of participants to find out whether it is perceived to be low in conscientiousness and/or low in agreeableness.

Study 4 a

Method

Participants and design. We collected data from 40 participants (18 female, 1 no specified gender, M_{age} = 24.13, SD = 6.97). Participants were recruited on the university campus and compensated with CHF 3 (Swiss Francs; about the same in US$ at the time) and a small chocolate present. Participants were randomly assigned to one of two different stimulus sets (see materials and procedure). The image classification task and the subsequent demographic questionnaires, however, were the same for all participants.

Materials. To create the stimuli for the image classification task we used the Basel Face Model (Paysan, Knothe, Amberg, Romdhani, & Vetter, 2009; Walker & Vetter, 2016), a multidimensional statistical face space derived from 200 3D scans of real faces. The dimensions of this space describe the shape (e.g., length, roundishness) and color information (e.g., darkness, contrast) with maximum variability between the 200 faces. Every face can be represented as a point in this multidimensional space, describing its values on all dimensions of the face space.

Vectors pointing from one face to another describe the difference between the two faces (for details see Vetter & Walker, 2011).

Combining the logic of the classical reverse correlation approach (Mangini & Biederman, 2004) with the face space approach (Stolier, Hehman, Keller, Walker, & Freeman, 2018; Walker & Keller, 2018), we created pairs of faces by applying random noise (i.e., random vectors in the face space) to a base face (i.e., the average face of the Basel Face Model). For each stimulus set, we created 98 vectors randomly manipulating shape information in faces, and 98 vectors randomly manipulating color information. Each random vector was once added to and once subtracted from the base face, resulting in 196 face pairs per set (i.e., 98 face pairs varying only in shape and 98 pairs varying only in color information; see *Figure 3* for two exemplar pairs).

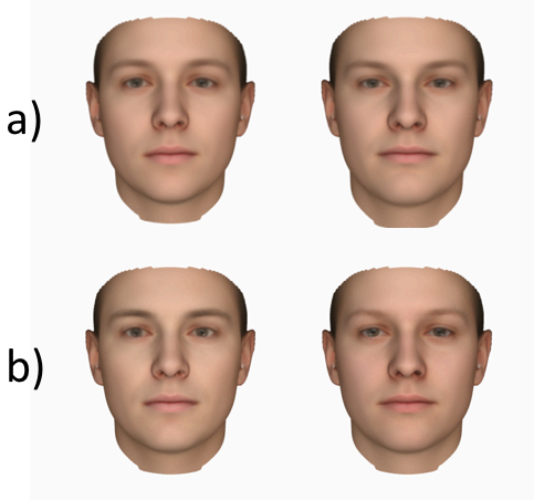


Figure 3. Two exemplar face pairs as presented in the image classification trials in Study 4a visualizing a) two faces only varying regarding shape information and b) two faces only varying regarding color information.

Procedure. We told participants before the task that there were many reasons to exclude another person and that there were no right or wrong answers. Participants were then asked to indicate as spontaneously as possible which one of two persons they would rather exclude from a group, by pressing one of two keys on the keyboard. In each trial, participants were presented with two faces, aligned horizontally on the screen. Each participant was presented with one of the

two stimulus sets, that is, 196 trials in total. In the first 98 trials, the two faces presented only varied regarding shape information, and in the second 98 trials, the two faces presented only varied regarding texture information. Trial order within the shape as well as within the texture trials was random. After 49 trials there was a short break and participants were free to continue whenever they were ready. After finishing the image classification task, participants answered a German version of the justice sensitivity scale (Schmitt, Baumert, Gollwitzer, & Maes, 2010). The scale was assessed for exploratory reasons unrelated to the present research question and will not be discussed further here. After finishing the questionnaire, participants were thanked and compensated.

Results

We calculated a vector indicating the mental representation of an ostracizable person's face. This was done by averaging all vectors underlying the faces that were selected for exclusion by participants, resulting in a single ostracism vector (for more details, see Keller, Reutner, Greifeneder, & Walker, 2018; Stolier et al., 2018; Walker & Keller, 2018). This combined ostracism vector was added to and subtracted from the base face, again to create two different visualizations. Adding the vector to the base face visualizes the mental representation of a person that individuals would rather ostracize (henceforth referred to as the *ostracism facial stereotype*). Subtracting the same vector from the base face visualizes the mental representation of a person that individuals would rather not ostracize (henceforth referred to as the *anti-ostracism facial stereotype;* additional studies from our lab that are unrelated to the present research question found that this mental representation is also equivalent to the face of a person one would actively include). See *Figure 4* for the visual results.

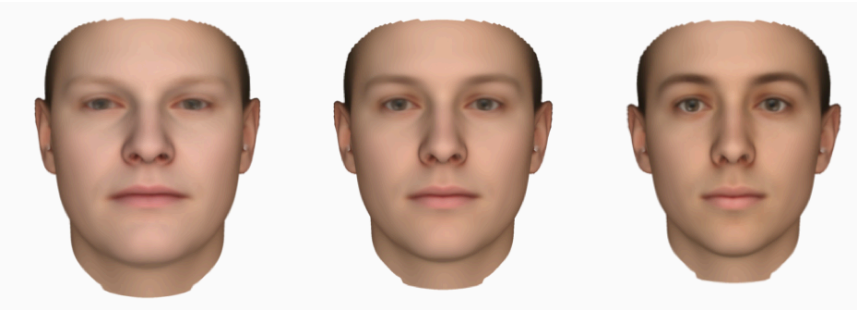


Figure 4. Visualization of the Ostracism facial stereotype (left), the average face from the Basel Face Model (middle), and the Anti-Ostracism facial stereotype (i.e., the face pointing in the opposite direction from the average face; right) as extracted in Study 4a.

Study 4 b

Study 4b investigates whether the effect of perceived conscientiousness and agreeableness on ostracism found in Studies 1-3 will emerge without directly manipulating said dimensions, but instead with a more subtle measure, namely impression formation from faces. To do so, we tested whether the stereotypical face of a person individuals would ostracize is perceived as low in conscientiousness and agreeableness. Thus, we presented both the ostracism and the anti-ostracism facial stereotype created in Study 4a to a new sample of participants and had them rate the two faces on the Big Five personality dimensions, as well as various other measures detailed below. We hypothesized that the ostracism facial stereotype would be rated as lower in conscientiousness and agreeableness than the anti-ostracism facial stereotype.

Method

Participants and design. We calculated the sample size to detect a medium-sized main effect of the presented stimulus faces on participants' rating of their personality (d = .50, power = .90, required n = 44). We thus recruited 52 students (27 female, M_{age} = 24.17, SD = 4.43) in the university cafeteria, who participated in the study for CHF 3.50 (about the same in US$ at the time). The study used a within-subject design and was preregistered on AsPredicted.org, see http://aspredicted.org/blind.php?x=ct857k

Materials and procedure. Participants were presented with the two faces generated in Study 4a, the stereotypical facial representation of an ostracized person (the ostracism facial stereotype) and the stereotypical facial representation of a person not to be ostracized (the anti-ostracism facial stereotype), in a random order. For each stimulus face, they were asked to indicate the spontaneous impression they had of that person. More specifically, they were asked to rate the depicted persons on the Big Five Personality Dimensions, using two items with highest factor loading from the BFI-K (Rammstedt & John, 2005; see Walker & Vetter, 2016). Thus, each dimension was tested by two items (e.g., reversed Agreeableness: "The presented person can be cold and distanced"). The stimulus faces were further rated on the Big Two (Agency and Communion, Runge, Frey, Gollwitzer, Helmreich, & Spence, 1981; the two items with the highest factors loadings were used, see Walker & Vetter, 2016), as well as attractiveness, sympathy, trustworthiness, and dominance. Moreover, we asked with two items each whether the person would comply with social norms, appeared familiar, and psychologically as well as physically healthy. All ratings were made on 7-point Likert scale *(1 = completely disagree; 7 = completely agree)*.

Results

A 2 (stimulus face: ostracism facial stereotype vs. anti-ostracism facial stereotype) x 5 (personality dimension: Conscientiousness vs. Agreeableness vs. Neuroticism vs. Openness vs. Extraversion) within-subject ANOVA revealed significant main effects for the stimulus face, $F(1,51) = 58.02$, $p < .001$, $\eta^2 = .53$, and the personality dimension, $F(4,48) = 24.11$, $p < .001$, $\eta^2 = .67$, that were both qualified by the significant stimulus face x personality dimension interaction, $F(4,48) = 10.05$, $p < .001$, $\eta^2 = .46$, such that the type of face affected some personality ratings more than others. Bonferroni-corrected simple main effects revealed that the ostracism facial stereotype was evaluated to be less conscientious, $p = .001$, $d = .51$ ($M = 3.78$, $SD = 1.15$ vs. $M =$

4.52, *SD* = 1.22), less agreeable, $p < .001$, $d = 1.10$ (*M* = 2.77, *SD* = 1.27 vs. *M* = 4.68, *SD* = 1.34), and less open to experience, $p < .001$, $d = .74$ (*M* = 2.81, *SD* = 1.21 vs. *M* = 4.26, *SD* = 1.46) compared to the anti-ostracism facial stereotype. Ostracism manipulations in the face did not significantly affect ascriptions of extraversion, $p = .086$, $d = .24$ (*M* = 3.48, *SD* = 1.29 vs. *M* = 3.95, *SD* = 1.25) and neuroticism, $p = .452$, $d = .14$ (*M* = 4.65, *SD* = 1.21 vs. *M* = 4.46, *SD* = 1.23).

As for the other measures, the stimulus persons were rated similar in agency, $t(51) = -1.78$, $p = .081$, $d = .25$ (*M* = 4.75, *SD* = 1.32 vs. *M* = 4.28, *SD* = 1.13), but the ostracism facial stereotype was evaluated to be lower in communion, $t(51) = 7.60$, $p < .001$, $d = 1.08$ (*M* = 3.0, *SD* = .90 vs. *M* = 4.64, *SD* = 1.33). The ostracism facial stereotype was also rated as less sympathetic, $t(51) = 8.90$, $p < .001$, $d = 1.24$ (*M* = 3.02, *SD* = 1.04 vs. *M* = 5.02, *SD* = 1.26), less attractive, $t(51) = 6.25$, $p < .001$, $d = .88$ (*M* = 2.60, *SD* = 1.24 vs. *M* = 4.08, *SD* = 1.63), less trustworthy, $t(51) = 7.36$, $p < .001$, $d = 1.02$ (*M* = 2.98, *SD* = 1.20 vs. *M* = 4.75, *SD* = 1.22), but more dominant, $t(51) = -4.57$, $p < .001$, $d = .63$ (*M* = 4.77, *SD* = 1.45 vs. *M* = 3.37, *SD* = 1.39) compared to the anti-ostracism facial stereotype. In addition, the ostracism facial stereotype was rated as less likely to comply with social norms, $t(51) = 6.14$, $p < .001$, $d = .85$ (*M* = 3.62, *SD* = 1.19 vs. *M* = 5.02, *SD* = 1.16), less familiar, $t(51) = 5.08$, $p < .001$, $d = .71$ (*M* = 2.65, *SD* = 1.47 vs. *M* = 3.94, *SD* = 1.61), less similar to the participants, $t(51) = 4.10$, $p < .001$, $d = .57$ (*M* = 2.00, *SD* = 1.10 vs. *M* = 2.96, *SD* = 1.39), and less likely to be physically and psychologically healthy, $t(51) = 4.67$, $p < .001$, $d = .66$ (*M* = 4.62, *SD* = 1.38 vs. *M* = 5.50, *SD* = .98) than the anti-ostracism facial stereotype.

Discussion

Study 4 tested the link between personality and ostracism intentions with a less direct paradigm. In a first part (Study 4a) using an intuitive, non-deliberate forced choice image

classification task, we generated an ostracism facial stereotype and its anti-face from vectors in the Basel Face Model (Paysan et al., 2009; Walker & Vetter, 2016). In the second part, these two faces were localized on the Big Five personality dimensions and additional measures (e.g., the Big Two personality dimensions) by a different sample of participants. As expected, the ostracism facial stereotype was rated as less conscientious and less agreeable than its anti-face. In addition, the person with the ostracizable face was also rated as less open, but similar with regard to extraversion and neuroticism. The findings indicate that individuals not only associate low conscientiousness and/or agreeableness with the likelihood of being ostracized, but that there is further a socially shared notion of which personality variables appear in the face of a person likely to be ostracized, namely low conscientiousness and agreeableness. Importantly, the results of the study attest to the robustness of our findings that also replicate on a less explicit measure as in Studies 1-3.

Study 5

Studies 1-4 present experimental evidence that individuals with certain personality characteristics are more likely ostracized, as well as demonstrate that there is a common facial conception or stereotype of an ostracized person that is linked to certain personality characteristics. However, a crucial question is whether such intentions are translated to the real life and whether the obtained results are thus ecologically valid. Assuming that the sources' intentions translate into behavior, and taking into account that individuals are highly sensitive for the experience of ostracism (Rudert, Hales, et al., 2017; Williams, 2009), one should expect that individuals who are less conscientious or agreeable report experiencing ostracism more often. It should be noted that we are not first to address this important question. Previous studies on the ostracism/personality link, which focused on the workplace and used small and specific samples (Nielsen et al., 2017; Wu et al., 2011), have found relations between self-reported ostracism and

conscientiousness, agreeableness, neuroticism, as well as openness. We here complement these findings by testing the hypothesized relations in a general context with a large-scale, representative sample from the German Socio-Economic panel (SOEP). Despite its importance, such evidence is lacking in the literature so far.

Method

Sample. We used data from the 2015 wave of the SOEP, a representative longitudinal survey of German households with almost 20 000 participants who are surveyed on an annual basis. Specifically, the scales that are relevant for our research question were part of the survey given to the innovation sample (SOEP-IS), a subsample of the SOEP that specifically allows for testing new research questions. In sum, 2745 individuals (53 % female, M_{age}= 52.44, *SD* = 18.26, *Range* $_{Year\ of\ Birth}$ = 1919 - 1998) answered the newly developed Ostracism Short Scale (OSS). Participants were nested within 1718 households. Of these, 881 households (51.3 %) provided one participant, 695 households (40.5 %) provided two participants, and 142 households (8.2 %) provided three or more participants.

Measures. ***The Ostracism Short Scale (OSS).*** To be able to contribute items to the nationwide representative SOEP-IS survey, we needed to provide a measurement for the frequency of ostracism that consists of very few items. To this end, we developed the Ostracism Short Scale (OSS) that was based on the Ferris scale (Ferris, Brown, Berry, & Lian, 2008). The OSS is a four-item scale measuring the general subjective frequency that a person had felt ostracized within the previous two months. In particular, participants were asked (in German): *"How often did you experience the following occurrences during the last two months?"* with respect to: *"Others ignored me," "Others shut me out from the conversation," "Others treated me as if I wasn't there," "Others did not invite me to activities."* All items are rated on a 7-point Likert scale *(1 = never, 7 = always).* We pretested the OSS in a sample with 174 participants (74

% female, M_{age}= 26.37, *SD* = 7.61, Range = 18 – 79, recruitment via the online pool of the German-speaking social psychology project "Forschung Erleben"). The OSS showed a high reliability (Cronbach's α = .87; compared to e.g., α = .89 of the original 10-item scale; see Ferris et al., 2008). Moreover, we tested for convergent validity by correlating the OSS with several other measures such as the Relatedness subscale from the Balanced Measure of Psychological Need Scale (Sheldon & Hilpert, 2012), and loneliness (Hawkley, Duvoisin, Ackva, Murdoch, & Luhmann, 2015; Luhmann & Hawkley, 2016). Criterion validity was achieved by correlating the OSS with the Satisfaction with Life Scale (Glaesmer, Grande, Braehler, & Roth, 2011), the World Health Organization-5 Well-being index (Brähler, Mühlan, Albani, & Schmidt, 2007), Need threat and mood (Rudert & Greifeneder, 2016), as well as life satisfaction measures (Lang, Weiss, Gerstorf, & Wagner, 2013; Schimmack, Krause, Wagner, & Schupp, 2010). The OSS showed medium correlations (highest *r* = .56 with need threat and loneliness, lowest *r* = -.33 for life satisfaction in the next five years), see *Table 6,* which speaks for the validity of the scale.

Table 6

Validity of the Ostracism Short Scale

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) Ostracism Short Scale									
(2) Life Satisfaction	**-.49****								
(3) Life Satisfaction (SOEP)	**-.47****	.78**							
(4) Life Satisfaction 1 y (SOEP)	**-.44****	.67**	.80**						
(5) Life Satisfaction 5 y (SOEP)	**-.33****	.65**	.71**	.86**					
(6) Need Threat	**.56****	-.62**	-.69**	-.63**	-.55**				
(7) Mood	**-.51****	.68**	.77**	.66**	.62**	-.89**			
(8) Well-being	**-.41****	.56**	.70**	.50**	.48**	-.57**	.63**		
(9) Relatedness	**-.47****	.62**	.57**	.44**	.48**	-.60**	.55**	.52**	
(10) Loneliness	**.56****	-.60**	-.54**	-.47**	-.45**	.62**	-.59**	-.47**	-.64**

Note. ** p < .001

The Big Five. Within the SOEP, the Big Five were assessed with the short scale BFI-S (Gerlitz & Schupp, 2005), which measures each of the five personality dimensions with three items using a seven-point Likert scale *(1 = completely disagree, 7 = completely agree*).

Results

The zero-order correlations revealed that ostracism was negatively correlated with conscientiousness, $r = -.20$, $p < .001$, agreeableness, $r = -.18$, $p < .001$, extraversion, $r = -.19$, $p < .001$, and positively with neuroticism, $r = .20$, $p < .001$. Openness was not significantly related to ostracism, $r = .02$, $p = .445$. When we regressed ostracism on all five predictors simultaneously, we found negative relations with conscientiousness, $\beta = -.13$, $p < .001$, agreeableness, $\beta = -.13$, $p < .001$, extraversion, $\beta = -.17$, $p < .001$, and a positive relation with neuroticism, $\beta = .16$, $p < .001$. Interestingly, now there emerged a *positive* relation between ostracism and openness, $\beta = .08$, $p < .001$. The variance of ostracism explained by the Big Five was $R^2 = .11$.

To test for temporal stability, we performed a regression predicting ostracism reported in 2015 from the Big Five reported in 2013. Again, we found negative relations with conscientiousness, $\beta = -.12$, $p < .001$, agreeableness, $\beta = -.07$, $p < .001$, extraversion, $\beta = -.12$, $p < .001$, and a positive relation with neuroticism, $\beta = -.06$, $p < .001$. There was no relation between ostracism and openness, $\beta = .01$, $p = .528$. The variance explained by the regression model was $R^2 = .06$.

Discussion

The data from the SOEP-IS showed evidence for a general link between the Big Five personality dimensions and experienced ostracism. Particularly, individuals who perceive themselves as more conscientious, agreeable, emotionally stable, and extraverted reported being ostracized less often. These relations could be found when we predicted ostracism experience based on data assessed at the same point in time as well as based on data collected two years

earlier, ruling out the specific measurement point at which participants took the survey as a possible explanation, and attesting to temporal stability. Given the representativeness of the sample, the data thus speaks to the generalizability of the obtained results, at least within Germany, a relatively culturally *independent* nation (Markus & Kitayama, 1991).

The present results partly dovetail with findings obtained by Wu and colleagues (2011). In particular, Wu and colleagues reported a survey with 208 employees in the petroleum industry in China, a relatively culturally *inter*dependent nation (Markus & Kitayama, 1991). They found negative relations of workplace ostracism with agreeableness and extraversion and a positive relation with neuroticism. The obtained relations for the experience of ostracism from the SOEP data are further in conceptual alignment with the meta-analysis of Nielsen and colleagues (2017), who investigated the effect of target personality on the experience of workplace harassment. Similar to our findings, the meta-analysis showed positive relations of experienced workplace harassment with conscientiousness, agreeableness, and extraversion, as well as a negative relation with neuroticism, but none with openness.

The SOEP data represents an important final link in our argumentation, as it indicates that the effects that we demonstrated within the lab transfer, at least partly, to the real world. One should note that the strength of the associations differs from what was observed in the experimental studies. Particularly, while conscientiousness and agreeableness again robustly predicted ostracism, other predictors proved to be equally strong. In contrast to the experimental studies, in the SOEP, openness was not a significant predictor, but extraversion was. One possible explanation is that due to the cross-sectional nature of the data, the obtained associations might be affected by several other processes (for example, highly open individuals may seek out more situations, including some where they encounter more ostracism). Moreover, there might be

reversed causality processes, that is, ostracism affecting a target's personality. We discuss this in more detail within the General Discussion.

General Discussion

Ostracism can be due to a variety of causes, one being the personality characteristics of the ostracized target. The present manuscript focusses on one particular mechanism, namely the target provoking ostracism because of his/her personality characteristics. We predicted and pre-registered our hypotheses that especially low agreeableness and low conscientiousness would elicit ostracism and found evidence for this within five studies and across three different paradigms. Studies 1-3 show that individuals report more ostracism intentions for targets that are described as disagreeable or careless. The effect was partially mediated by liking (Studies 2 and 3), but not additive, given that being described as negative on one personality dimension could not be compensated by being described as positive on the other (Study 3). We further found that there is a socially shared facial stereotype of how a person likely to be ostracized appears, which is that of a careless and disagreeable person (Study 4). Finally, when analyzing data from a representative survey (the SOEP-IS), self-reported agreeableness and conscientiousness reliably predicted experienced ostracism, even when ostracism was measured two years after personality.

The obtained results pose an important step in understanding the reasons for why ostracism occurs in the first place, a question that has played a minor role in ostracism research so far. We believe this questions to be a highly important one, given that to prevent ostracism, one needs to know first from where it derives. Of course, there are a variety of reasons why ostracism may occur, many of them are due to situational circumstances or even purely incidental because of random selection mechanisms (Lindström & Tobler, 2018). Moreover, different causes may operate at different levels of explanation; a target's personality is a relatively *ultimate*, or upstream cause of later ostracism, while a source's disliking - elicited by the target's

personality- represents a more *proximate*, or downstream cause. As personality can be understood as accumulated behavior of a person over time (e.g., Fleeson, 2001), it is likely at least one crucial reason for why some individuals are more likely to become targets of ostracism. While the experimental studies (Studies 1-4) shed light on one of the potential mechanisms by which personality might increase ostracism intentions of the sources, Study 5 demonstrates that the respective relations also exist in real life and can be shown in a representative sample. Together with the results from non-representative survey studies in other countries, this speaks for the generalizability and stability of our results across cultures and standard demographic variables.

The Source Perspective: Motivations for Ostracism

Within the scope of this manuscript, we mainly focused on the sources' perspective (Studies 1-4) and proposed two reasons why individuals would ostracize others: First, the target has violated social norms, which we assumed to be the case for individuals perceived as disagreeable, and second, the target represents a burden to the group, which we assumed to be the case for individuals perceived as low in conscientiousness. It should be emphasized that these reasons are not necessarily mutually exclusive, in fact, it is plausible that in some situations, they might even influence one another. For instance, a target that constantly violates group norms is likely to be perceived burdensome at some point. Inversely, a continuous underperformance might be interpreted as social loafing and a violation of a group's performance norms. This finding is also in line with the "negative halo effect" demonstrated in Study 3, showing that an individual described as low in conscientiousness was also perceived as less agreeable and vice versa. Rather than representing a methodological caveat, we believe that it is highly probable that individuals' assumptions of how certain personality dimensions are related are in line with actual correlations between the respective personality dimensions (Soto & John, 2017).

Effects of extreme personality dispositions

Studies 1 and 2 show that ostracism intentions mainly increase due to individuals having negative personality dispositions, such as being disagreeable or careless. Being positive on one personality dimension did not decrease ostracism intentions further compared to the control group. One can speculate, though, that a person with an extreme personality disposition in a positive direction could also face an increased risk of ostracism. For instance, under certain situational circumstances, a highly conscientious overperformer can get bothersome and pose a threat to a group, as s/he may make the rest of the group look less competent or threaten the position or status of single group members (Maner & Mead, 2010). Moreover, obsessive-compulsive personality disorder (OCPD) has been described as an extreme, maladaptive form of conscientiousness (Samuel & Widiger, 2011) that is characterized by perfectionism and an unhealthy preoccupation with order and organization. It is easy to see how such a person might become a burden for a group as well, and thus be at an increased risk of becoming a target of ostracism. This assumption is also in line with research showing that individuals dislike others the more they perceive these others to have extreme dispositions (Koch, Imhoff, Dotsch, Unkelbach, & Alves, 2016). Given that dislike partly mediated the link between personality and ostracism intentions, it is plausible to assume that being extreme on a specific personality disposition might increase the risk of becoming a target of ostracism, too. However, such a potential negative effect of an extreme personality would probably only hold for some dispositions but not for others; in particular, it is less easy to imagine that a person could be ostracized due to being too agreeable, given the centrality of warmth in person perception (Koch et al., 2016).

Neuroticism, Openness, and Extraversion

This contribution focusses on the effects of conscientiousness and agreeableness, as it was our impression that the existing literature allows for rather clear-cut predictions for these two traits. This impression was supported by the data collected. Despite this (pre-registered) focus, in most studies (with exception of Study 3), we also assessed the remaining Big Five Personality Dimensions in an exploratory fashion. Although speculative and less clear, we would like to briefly discuss the different patterns we found for Neuroticism, Openness, and Extraversion, and offer some potential explanations that might become the basis for confirmatory research on these relations.

Perceived *Neuroticism* increased ostracism intentions in Studies 1 and 2, moreover, self-reported neuroticism in the SOEP-IS was positively related to self-reported ostracism. Interestingly, in Study 4, neuroticism did not seem to be part of the facial stereotype of how a person, which individuals would ostracize, looks like. One possible explanation for this is that in real life, individuals generally cannot reliably detect neuroticism (Vazire, 2010) and thus, neuroticism may not be a part of the facial ostracism stereotype. Alternatively, while highly anxious and nervous individuals might both be more sensitive to social ostracism as well as perceived as bothersome and thus be ostracized more often, there might be less of a social consensus to do so. As previous research has demonstrated, individuals are aware that ostracism is painful for the ostracized person (Legate et al., 2013; Wesselmann et al., 2009) and strongly disapprove of groups ostracizing individuals that can possibly not take care of themselves and depend on the protection of the group (Rudert, Reutner, et al., 2017). Thus, while ostracizing an emotionally unstable person may come with benefits for the group, it also carries the risk of social (disapproval) costs. Further research may fruitfully investigate whether the here obtained (facial) finding generalizes to explicitly expressed stereotypes, too.

Low *Openness* increased ostracism intentions in Studies 1 and 2 and was also part of the facial ostracism stereotype in Study 4. However, in the representative data from the SOEP-IS (Study 5), self-reported openness was not significantly, or even positively, related to self-reported ostracism. We can think of two potential explanations for this finding: First, when asked directly, individuals might over-estimate the positivity of high openness. While typically perceived as a positive characteristic, openness is also linked to unconventionality and a tendency to disregard social norms (McCrae & Sutin, 2009), which could be a reason for others to ostracize a very open individual. Second, while others might be more inclined to ostracize close-minded persons, close-minded persons may generally not approach new situations very often. As a consequence, close-minded individuals might be less likely to encounter situations in which they might experience ostracism. Individuals high in openness, instead, might seek unfamiliar environments that come with the risk of rejection more often (for example, parties populated mostly with peripheral social connections). These ecological effects may be reflected in the survey but not in our laboratory studies, thus offering one explanation for the difference between the negative effects we found in the experiments and the null, or even positive effect in the representative sample.

Finally, low *Extraversion* did not predict ostracism in any of the experiments, nor was Extraversion a part of the facial ostracism stereotype in Study 4. However, we found a substantial and stable negative relation between self-reported Extraversion and self-reported ostracism in Study 5. A plausible explanation for this finding would be *oblivious ostracism*: introverted people might be overlooked more often and consequentially experience ostracism more often, even though the involuntary sources of ostracism might not have intentionally planned to ostracize them (Lindström & Tobler, 2018; Williams, 1997).

Methodological Considerations

It should be noted that while all studies within the present contribution investigate the relation between ostracism and personality, they differ in their focus and the investigated perspective: Studies 1-3 investigate the motivation of the sources and their intention to ostracize others, with personality being manipulated and presented to participants. In daily life, people can detect certain traits more accurately than others. For example, extraversion, characterized by sociability and visible behaviors, tends to be accurately detected, while neuroticism, characterized by distressed internal mental states tends to be less accurately detected (Vazire, 2010). It follows then, that in real life, ostracism intentions might further be affected by how well individuals can detect certain traits, such as there could be stronger biases for traits which individuals can detect less well. Alternatively, it is also possible that individuals might be aware of their ability to detect certain traits and will be more careful to ostracize persons due to personality dimensions they can detect less well, as we have previously discussed for neuroticism.

Study 4 focusses on the stereotypical perception of individuals who become a target of ostracism. Although closely related to Studies 1-3, this study uses a very subtle measurement and thus demonstrates that the suggested association between personality and ostracism is present even if individuals cannot easily discern what is measured and without directly manipulating the personality dimensions of interest.

In contrast to the other studies, Study 5 focuses on the perceptions and self-evaluation of the targets of ostracism. While agreeableness and conscientiousness were again related to ostracism, it appears likely that in this study, additional mechanisms aside from the motivation of the sources influence this relation. This assumption is also in line with differences in the detected patterns: In the experimental studies, conscientiousness and agreeableness were the strongest

predictors of ostracism intentions, while in the SOEP-IS data, extraversion and neuroticism predicted ostracism more strongly. It is thus likely that the effects of extraversion and neuroticism are mainly due to other mechanisms than the target provoking the ostracism.

Two of these alternative processes may reflect influences of *target perception*, that is, individuals with certain personality dispositions interpreting negative events as ostracism, and *reversed causality*, that is, individual's personality changing as an effect of being repeatedly being ostracized (Nielsen et al., 2017). In real life, it is highly likely that the three mechanisms (target provocation, target perception, and reversed causality) are strongly interconnected in vicious circles: For instance, individuals who tend to interpret even minor incidents as intentional ostracism by others, may as a result behave in manners that provoke actual ostracism, which would then confirm individuals in their beliefs. Respective processes have been shown for rejection sensitivity (Downey et al., 1998; Downey et al., 2004), a disposition that is closely related to neuroticism, and for agreeableness (Hales et al., 2016).

Conclusions

In five studies, across three different paradigms, with participants from different cultural backgrounds, and with one nation-wide representative sample, we investigated the link between personality dispositions and ostracism. We find that especially low agreeableness and low conscientiousness were reliably associated with a higher risk of becoming a target of ostracism. The presented findings enhance our understanding why and under which conditions ostracism occurs in the first place, which is important to know when aiming at its prevention.

References

Ashton, M. C., Lee, K., & Paunonen, S. V. (2002). What is the central feature of extraversion? Social attention versus reward sensitivity. *Journal of Personality and Social Psychology, 83*(1), 245 - 252. doi:10.1037/0022-3514.83.1.245

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1-26. doi:10.1111/j.1744-6570.1991.tb00688.x

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology, 92*(2), 410 - 424. doi:10.1037/0021-9010.92.2.410

Brähler, E., Mühlan, H., Albani, C., & Schmidt, S. (2007). Teststatistische Prüfung und Normierung der deutschen Versionen des EUROHIS-QOL Lebensqualität-Index und des WHO-5 Wohlbefindens-Index. *Diagnostica, 53*(2), 83-96. doi:10.1026/0012-1924.53.2.83

Cain, S. (2013). *Quiet: The power of introverts in a world that cant stop talking*. London: Penguin Books.

Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13*(6), 653-665. doi:10.1016/0191-8869(92)90236-I

Coyne, S. M., Nelson, D. A., Robinson, S. L., & Gundersen, N. C. (2011). Is viewing ostracism on television distressing? *The Journal of Social Psychology, 151*(3), 213-217. doi:10.1080/00224540903365570

Ditrich, L., & Sassenberg, K. (2016). It's either you or me! Impact of deviations on social exclusion and leaving. *Group Processes & Intergroup Relations, 19*(5), 630-652. doi:10.1177/1368430216638533

Downey, G., Freitas, A. L., Michaelis, B., & Khouri, H. (1998). The self-fulfilling prophecy in close relationships: Rejection sensitivity and rejection by romantic partners. *Journal of Personality and Social Psychology, 75*(2), 545-560. doi:10.1037/0022-3514.75.2.545

Downey, G., Mougios, V., Ayduk, O., London, B. E., & Shoda, Y. (2004). Rejection sensitivity and the defensive motivational system: Insights from the startle response to rejection cues. *Psychological Science, 15*(10), 668-673. doi:10.1111/j.0956-7976.2004.00738.x

Ferris, D. L., Brown, D. J., Berry, J. W., & Lian, H. (2008). The development and validation of the Workplace Ostracism Scale. *Journal of Applied Psychology, 93*(6), 1348-1366. doi:10.1037/a0012743

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology, 38*(6), 889-906. doi:10.1037/0022-3514.38.6.889

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77-83. doi:10.1016/j.tics.2006.11.005

Fleeson, W. (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology, 80*(6), 1011 -1027. doi:10.1037/0022-3514.80.6.1011

Funk, F., Walker, M., & Todorov, A. (2017). Modelling perceptions of criminality and remorse from faces using a data-driven computational approach. *Cognition and Emotion, 31*(7), 1431-1443. doi:10.1080/02699931.2016.1227305

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., . . . Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-nation study. *Science, 332*(6033), 1100-1104. doi:10.1126/science.1197754

Gerlitz, J.-Y., & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. *DIW Research Notes, 4*.

Giner-Sorolla, R. (2018). Powering our interaction. Retrieved from https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/

Glaesmer, H., Grande, G., Braehler, E., & Roth, M. (2011). The German version of the satisfaction with life scale (SWLS). *European Journal of Psychological Assessment, 27*(2), 127-132. doi:10.1027/1015-5759/a000058

Graziano, W. G., & Eisenberg, N. (1997). Agreeableness: A dimension of personality. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of Personality Psychology* (pp. 795-824). Cambridge, MA: Academic Press.

Graziano, W. G., Habashi, M. M., Sheese, B. E., & Tobin, R. M. (2007). Agreeableness, empathy, and helping: A person× situation perspective. *Journal of Personality and Social Psychology, 93*(4), 583-599. doi:0.1037/0022-3514.93.4.583

Güroğlu, B., Will, G.-J., & Klapwijk, E. T. (2013). Some bullies are more equal than others: Peer relationships modulate altruistic punishment of bullies after observing ostracism. *International Journal of Developmental Science, 7*(1), 13-23. doi:10.3233/DEV-1312117

Hales, A. H., Kassner, M. P., Williams, K. D., & Graziano, W. G. (2016). Disagreeableness as a cause and consequence of ostracism. *Personality and Social Psychology Bulletin, 42*(6), 782-797. doi:10.1177/0146167216643933

Hawkley, L. C., Duvoisin, R., Ackva, J., Murdoch, J. C., & Luhmann, M. (2015). *Loneliness in older adults in the USA and Germany: Measurement invariance and validation*. Retrieved from http://www.norc.org/PDFs/Working%20Paper%20Series/WP-2015-004.pdf

Hayes, A. F. (2013). *Introduction to mediation, moderation and conditional process analysis.* New York, NY: Guilford.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*(6), 869-875. doi:10.1037/0021-9010.85.6.869

Jensen-Campbell, L. A., Adams, R., Perry, D. G., Workman, K. A., Furdella, J. Q., & Egan, S. K. (2002). Agreeableness, extraversion, and peer relations in early adolescence: Winning friends and deflecting aggression. *Journal of Research in Personality, 36*(3), 224-251. doi:10.1006/jrpe.2002.2348

Kagel, J., & McGee, P. (2014). Personality and cooperation in finitely repeated prisoner's dilemma games. *Economics Letters, 124*(2), 274-277. doi:10.1016/j.econlet.2014.05.034

Keller, M. D., Reutner, L., Greifeneder, R., & Walker, M. (2018). *Random faces telling stories: Combining a face space approach with reverse correlation to visualize internal representations*. Paper presented at the Annual Meeting of the Society for Personality and Social Psychology (SPSP), Atlanta, USA.

Kerr, N. L., & Levine, J. M. (2008). The detection of social exclusion: Evolution and beyond. *Group Dynamics: Theory, Research, and Practice, 12*(1), 39-52. doi:10.1037/1089-2699.12.1.39

Kervyn, N., Yzerbyt, V. Y., Demoulin, S., & Judd, C. M. (2008). Competence and warmth in context: The compensatory nature of stereotypic views of national groups. *European Journal of Social Psychology, 38*(7), 1175-1183. doi:10.1002/ejsp.526

Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology, 110*(5), 675-709. doi:10.1037/pspa0000046

Kurzban, R., & Leary, M. R. (2001). Evolutionary origins of stigmatization: The functions of social exclusion. *Psychological Bulletin, 127*(2), 187-208. doi:10.1037/0033-2909.127.2.187

Lang, F. R., Weiss, D., Gerstorf, D., & Wagner, G. G. (2013). Forecasting life satisfaction across adulthood: Benefits of seeing a dark future? *Psychology and Aging, 28*(1), 249-261. doi:10.1037/a0030797

Legate, N., DeHaan, C. R., Weinstein, N., & Ryan, R. M. (2013). Hurting you hurts me too: The psychological costs of complying with ostracism. *Psychological Science, 24*(4), 583-588. doi:10.1177/0956797612457951

Leising, D., Erbs, J., & Fritz, U. (2010). The letter of recommendation effect in informant ratings of personality. *Journal of Personality and Social Psychology, 98*(4), 668-682. doi:10.1037/a0018771

Lindström, B., & Tobler, P. N. (2018). Incidental ostracism emerges from simple learning mechanisms. *Nature Human Behaviour*, 405–414. doi:10.1038/s41562-018-0355-y

Luhmann, M., & Hawkley, L. C. (2016). Age differences in loneliness from late adolescence to oldest old age. *Developmental Psychology, 52*(6), 943. doi:10.1037/dev0000117

Maner, J. K., & Mead, N. L. (2010). The essential tension between leadership and power: When leaders sacrifice group goals for the sake of self-interest. *Journal of Personality and Social Psychology, 99*(3), 482-497. doi:10.1037/a0018559

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science, 28*(2), 209-226. doi:10.1207/s15516709cog2802_4

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*(2), 224 - 253. doi:10.1037//0033-295x.98.2.224

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality, 60*(2), 175-215. doi:10.1111/j.1467-6494.1992.tb00970.x

McCrae, R. R., & Sutin, A. R. (2009). Openness to experience. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of Individual Differences in Social Behavior* (Vol. 15, pp. 257-273). New York, NY: Guilford Publications.

McGrath, J. E. (1984). *Groups: Interaction and performance* (Vol. 14): Prentice-Hall Englewood Cliffs, NJ.

Milam, A. C., Spitzmueller, C., & Penney, L. M. (2009). Investigating individual differences among targets of workplace incivility. *Journal of Occupational Health Psychology, 14*(1), 58-69. doi:10.1037/a0012683

Nezlek, J. B., Wesselmann, E. D., Wheeler, L., & Williams, K. D. (2012). Ostracism in everyday life. *Group Dynamics: Theory, Research, and Practice, 16*(2), 91-104. doi:10.1037/a0028029

Nezlek, J. B., Wesselmann, E. D., Wheeler, L., & Williams, K. D. (2015). Ostracism in everyday life: The effects of ostracism on those who ostracize. *The Journal of Social Psychology, 155*(5), 432-451. doi:10.1080/00224545.2015.1062351

Nielsen, M. B., Glasø, L., & Einarsen, S. (2017). Exposure to workplace harassment and the Five Factor Model of personality: A meta-analysis. *Personality and Individual Differences, 104*, 195-206. doi:10.1016/j.paid.2016.08.015

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*(4), 250-256. doi:10.1037/0022-3514.35.4.250

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences, 43*(5), 971-990. doi:10.1016/j.paid.2007.03.017

Olivola, C. Y., & Todorov, A. (2010). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology, 46*(2), 315-324. doi:10.1016/j.jesp.2009.12.002

Over, H., & Uskul, A. K. (2016). Culture moderates children's responses to ostracism situations. *Journal of Personality and Social Psychology, 110*(5), 710 - 724. doi:10.1037/pspi0000050

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 225-239). New York, NY: Guilford Press.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. *6th IEEE International Conference on Advanced Video and Signal based Surveillance for Security, Safety and Monitoring in Smart Environments, Italy*, 296 - 301.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322-338. doi:10.1037/a0014996

Rammstedt, B., & John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K). *Diagnostica, 51*(4), 195-206. doi:10.1026/0012-1924.51.4.195

Ren, D., Wesselmann, E., & Williams, K. D. (2016). Evidence for another response to ostracism: Solitude seeking. *Social Psychological and Personality Science, 7*(3), 204-212. doi:10.1177/1948550615616169

Rothmann, S., & Coetzer, E. P. (2003). The big five personality dimensions and job performance. *SA Journal of Industrial Psychology, 29*(1), 68-74. doi:10.4102/sajip.v29i1.88

Rudert, S. C., & Greifeneder, R. (2016). When it's okay that I don't play: Social norms and the situated construal of social exclusion. *Personality and Social Psychology Bulletin, 42*(7), 955-969. doi:10.1177/0146167216649606

Rudert, S. C., Hales, A. H., Greifeneder, R., & Williams, K. D. (2017). When silence is not golden: Why acknowledgement matters even when being excluded. *Personality and Social Psychology Bulletin, 43*(5), 678 - 692. doi:10.1177/0146167217695554

Rudert, S. C., Reutner, L., Greifeneder, R., & Walker, M. (2017). Faced with exclusion: Perceived facial warmth and competence influence moral judgments of social exclusion. *Journal of Experimental Social Psychology, 68*, 101-112. doi:10.1016/j.jesp.2016.06.005

Rudert, S. C., Ruf, S., & Greifeneder, R. (2018). Who's to punish? How observers sanction norm-violating behavior in ostracism situations. *Manuscript submitted for publication*.

Rudert, S. C., Sutter, D., Corrodi, C., & Greifeneder, R. (2018). Who's to blame? Dissimilarity as a cue in moral judgments of observed ostracism episodes. *Journal of Personality and Social Psychology, 115*, 31-53. doi:10.1037/pspa0000122

Runge, T. E., Frey, D., Gollwitzer, P. M., Helmreich, R. L., & Spence, J. T. (1981). Masculine (instrumental) and feminine (expressive) traits: A comparison between students in the United States and West Germany. *Journal of Cross-Cultural Psychology, 12*(2), 142-162. doi:10.1177/0022022181122002

Samuel, D. B., & Widiger, T. A. (2011). Conscientiousness and obsessive-compulsive personality disorder. *Personality Disorders: Theory, Research, Treatment, 2*(3), 161-174. doi:10.1037/a0021216

Scheepers, D., Branscombe, N. R., Spears, R., & Doosje, B. (2002). The emergence and effects of deviants in low and high status groups. *Journal of Experimental Social Psychology, 38*(6), 611-617. doi:10.1016/S0022-1031(02)00506-1

Schimmack, U., Krause, P., Wagner, G. G., & Schupp, J. (2010). Stability and change of well being: An experimentally enhanced latent state-trait-error analysis. *Social Indicators Research, 95*(1), 19-31. doi:10.1007/s11205-009-9443-8

Schmitt, M., Baumert, A., Gollwitzer, M., & Maes, J. (2010). The Justice Sensitivity Inventory: Factorial validity, location in the personality facet space, demographic pattern, and normative data. *Social Justice Research, 23*(2-3), 211-238. doi:10.1007/s11211-010-0115-2

Sheldon, K. M., & Hilpert, J. C. (2012). The balanced measure of psychological needs (BMPN) scale: An alternative domain general measure of need satisfaction. *Motivation and Emotion, 36*(4), 439-451. doi:10.1007/s11031-012-9279-4

Simonsohn, U. (Producer). (2014). No-way interactions. Retrieved from http://datacolada.org/17

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*(1), 131-142. doi:10.1037/0033-2909.105.1.131

Snyder, M., & Gangestad, S. (1982). Choosing social situations: Two investigations of self-monitoring processes. *Journal of Personality and Social Psychology, 43*(1), 123-135. doi:10.1037/0022-3514.43.1.123

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117-143. doi:10.1037/pspp0000096

Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1807222115

Twenge, J. M., Baumeister, R. F., DeWall, C. N., Ciarocco, N. J., & Bartels, J. M. (2007). Social exclusion decreases prosocial behavior. *Journal of Personality and Social Psychology, 92*(1), 56-66. doi:10.1037/0022-3514.92.1.56

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology, 98*(2), 281-300. doi:10.1037/a0017908

Vetter, T., & Walker, M. (2011). Computer-generated images in face perception. In A. Calder, J. V. Haxby, M. Johnson, & G. Rhodes (Eds.), *The Oxford Handbook of Face Perception* (pp. 388-399). Oxford, UK: Oxford University Press.

Walker, M., Jiang, F., Vetter, T., & Sczesny, S. (2011). Universals and cultural differences in forming personality trait judgments from faces. *Social Psychological and Personality Science, 2*(6), 609-617. doi:10.1177/1948550611402519

Walker, M., & Keller, M. D. (2018). Beyond attractiveness: A multi-method approach to study enhancement in self-recognition on the Big Two personality dimensions. *Manuscript in revision.*

Walker, M., Schönborn, S., Greifeneder, R., & Vetter, T. (2018). The Basel Face Database: A validated set of photographs reflecting systematic differences in Big Two and Big Five personality dimensions. *PloS one, 13*(3), e0193190. doi:10.1371/journal.pone.0193190

Walker, M., & Vetter, T. (2016). Changing the personality of a face: Perceived Big Two and Big Five personality factors modeled in real photographs. *Journal of Personality and Social Psychology, 110*(4), 609-624. doi:10.1037/pspp0000064

Wesselmann, E. D., Bagg, D., & Williams, K. D. (2009). "I feel your pain": The effects of observing ostracism on the ostracism detection system. *Journal of Experimental Social Psychology, 45*(6), 1308-1311. doi:10.1016/j.jesp.2009.08.003

Wesselmann, E. D., Williams, K. D., & Wirth, J. H. (2014). Ostracizing group members who can (or cannot) control being burdensome. *Human Ethology Bulletin, 29*(2), 82-103.

Wesselmann, E. D., Wirth, J. H., Pryor, J. B., Reeder, G. D., & Williams, K. D. (2013). When do we ostracize? *Social Psychological and Personality Science, 4*(1), 108-115. doi:10.1177/1948550612443386

Wesselmann, E. D., Wirth, J. H., Pryor, J. B., Reeder, G. D., & Williams, K. D. (2015). The role of burden and deviation in ostracizing others. *The Journal of Social Psychology, 155*(5), 483-496. doi:10.1080/00224545.2015.1060935

Will, G.-J., Crone, E. A., van den Bos, W., & Güroğlu, B. (2013). Acting on observed social exclusion: Developmental perspectives on punishment of excluders and compensation of victims. *Developmental Psychology, 49*(12), 2236-2244. doi:10.1037/a0032299

Williams, K. D. (1997). Social ostracism. In R. M. Kowalski (Ed.), *Aversive Interpersonal Behaviors. The Social/Clinical Psychology* (pp. 133-170). Boston, MA: Springer.

Williams, K. D. (2009). Ostracism: A temporal need-threat model. In P. Z. Mark (Ed.), *Advances in experimental social psychology* (Vol. 41, pp. 275-314). San Diego, CA: Elsevier Academic Press.

Wirth, J. H., Lynam, D. R., & Williams, K. D. (2010). When social pain is not automatic: Personality disorder traits buffer ostracism's immediate negative impact. *Journal of Research in Personality, 44*(3), 397-401. doi:10.1016/j.jrp.2010.03.001

Wu, L., Wei, L., & Hui, C. (2011). Dispositional antecedents and consequences of workplace ostracism: An empirical examination. *Frontiers of Business Research in China, 5*(1), 23-44. doi:10.1007/s11782-011-0119-2

Zadro, L., Boland, C., & Richardson, R. (2006). How long does it last? The persistence of the effects of ostracism in the socially anxious. *Journal of Experimental Social Psychology, 42*(5), 692-697. doi:10.1016/j.jesp.2005.10.007

Appendix B

Keller, M. D., Reutner, L., Greifeneder, R., & Walker, M. (2019). *Faces evoking emotions stereotypically triggered by groups: Developing an advanced reverse correlation technique.* Manuscript under review.

Running head: ADVANCED REVERSE CORRELATION

Faces Evoking Emotions Stereotypically Triggered by Groups: Developing an Advanced Reverse Correlation Technique

Matthias David Keller, Leonie Reutner, Rainer Greifeneder, & Mirella Walker

University of Basel

Author Note

Correspondence concerning the paper should be addressed to Matthias David Keller, Department of Social Psychology, University of Basel, Missionsstrasse 62/64, 4055 Basel, Switzerland. Email: matt.keller@unibas.ch, Phone: ++41 +61 207 06 14.

Abstract

People naturally and spontaneously infer many attributes from a person's face. Moreover, faces evoke emotions. However, what is it about a face that leads to emotional reactions such as pity? Does a face that evokes admiration look similar to a face that evokes envy or do they differ? Can a face that evokes pity also evoke disgust? We aim to answer these questions by presenting an advanced reverse correlation technique. Using this technique, we extracted the prototypes of faces that evoke admiration, envy, pity, disgust, and fear (Study 1). A successful validation of the extracted prototypes (Study 2), a replication with non-manipulated faces (Study 3), and a study focusing on conceptual similarity between emotions (Study 4) revealed an admiration-envy and a disgust-fear similarity. Finally, we mapped the emotion prototypes onto the two-dimensional space of warmth and competence as used by the Stereotype Content Model (Study 5). These findings suggest a shared social consensus of what faces triggering specific emotions look like, and that people have similar representations of someone who evokes admiration and of someone who evokes envy on the one hand, and of someone who evokes disgust and of someone who evokes fear on the other hand. We highlight that the here presented technique makes it possible to visualize prototypes in a realistic manner and might serve as a valuable instrument to create stimuli for future research.

Keywords*: reverse correlation, statistical face models, impression formation, emotions, stereotype content model*

Abstract: 228 words

Faces Evoking Emotions Stereotypically Triggered by Groups: Developing an Advanced Reverse Correlation Technique

With the rise of social media platforms such as Instagram or Tinder, it has become increasingly important to form quick impressions of people solely or primarily from the photos they post. Given the societal implications, there has been increasing interest in identifying how judgments are made from a mere glance at someone's face. Consequently, there is extensive knowledge on what attributes are ascribed to individuals based on their facial appearance and what kind of appearances evoke which kind of personality ascriptions. However, very little is known about what facial characteristics evoke specific emotions. Further, little is known about perceived similarities of and differences between faces that evoke specific emotions. Does a face that evokes admiration look similar to a face that evokes envy or do they differ? Can a face that evokes pity also evoke disgust? Given the great importance of emotional reactions in predicting subsequent judgements and behavior (e.g., Lerner & Keltner, 2000; Mauss, McCarter, Levenson, Wilhelm, & Gross, 2005; see Lench, Flores, & Bench, 2011 for a review), the central aim of the present paper is to fill this gap and to visualize what faces that evoke specific emotional reactions look like. To do so, we borrow the conceptual frame from literature on stereotyped groups, specifically the Stereotype Content Model (Fiske, Cuddy, Glick, & Xu, 2002), which suggests that different social groups are associated with specific stereotypes, that can be located in a two-dimensional space of warmth and competence and identifies specific emotional reactions towards the respective social groups. Moreover, literature on face perception points out two principal dimensions, trustworthiness and dominance, that describe face evaluation (Oosterhof & Todorov, 2009) that share some similarities with the dimensions of group perception. Thus, the question arises as to what someone who evokes a specific emotion looks like, even if no explicit information about group affiliation is available.

Against this background, the aim of the present contribution is to extract prototypes of faces that evoke emotions that are known to be evoked by social groups and to investigate the structure of how the different emotions are evoked by faces. In order to extract these prototypes, we present a significant advancement of the traditional reverse correlation technique (Kontsevich & Tyler, 2004; Mangini & Biederman, 2004) by combining it with a statistical face space (Paysan, Knothe, Amberg, Romdhani, & Vetter, 2009) and up-to-date computer graphics techniques (Walker & Vetter, 2016).

We start our literature review with research on the SCM and person perception. In a second step, we focus on different reverse correlation techniques and how we combine them to unite their advantages.

Group Perception

A large body of work on the content of stereotypes has established that warmth and competence capture the essence of stereotypes (SCM; Caprariello, Cuddy, & Fiske, 2009; Cuddy, Fiske, & Glick, 2007; Fiske, Cuddy, & Glick, 2007; Fiske et al., 2002; for a recent development of this framework see Koch, Imhoff, Dotsch, Unkelbach, & Alves, 2016). In their original work, Fiske and colleagues (2002) asked their participants in an open-ended questionnaire to write down the "various types of people [they thought] today's society categorizes into groups" (p. 884). In a second step the most frequently mentioned groups were rated on a variety of traits. A principal component analysis resulted in two factors that best explained the ascribed traits of the different groups. Those two factors were identified as competence (e.g., competent, intelligent, ambitious) and warmth (e.g., friendly, sincere, trustworthy). For instance, the SCM allows for the deductions that US-individuals associate the social group of American Football players with stereotypic traits such as athletic, ambitious, or sincere; that this group will be located in the quadrant of high competence and high warmth.

Arguing from a functional perspective, Fiske and colleagues (2002) suggest that if people meet groups or if people meet individuals from specific groups, they want to know what the other's intentions are (warmth) and whether the individual or the group is able to act upon these intentions (competence). As a result, four clusters of groups appear that are perceived as either high on both dimensions, low on both dimensions, or high on the one dimension and low on the other dimension (Fiske et al., 2002).

Building on the cognitive ascriptions of warmth and competence, the SCM further identifies specific emotional reactions. Groups perceived as warm and competent elicit feelings of admiration or pride (e.g., students, when asking a sample consisting of American students, or American Football players, as in the introductory example). Groups perceived as cold and incompetent elicit feelings of disgust or contempt (e.g., drug addicted). Groups perceived as warm but incompetent elicit feelings of pity (e.g., elderly people). Groups perceived as cold but competent elicit feelings of envy (e.g., rich people; Cuddy et al., 2007; Fiske et al., 2002).

Person Perception

Based on explicit group affiliation. Fiske and colleagues (2002) built their model on perceptions of groups. Applying this group-based model to the perception of individuals, one may similarly assume that the affiliation of an individual to a specific social group activates stereotype content, which seamlessly evokes specific emotions among perceivers. Therefore, one may deduce that providing explicit context information about a person's group affiliation may create a task that resembles the original task of imagining a specific group and reporting which emotions are evoked when thinking of this group (see for example Cuddy et al., 2007 or Fiske et al., 2002). To illustrate with the example of American Football players, let us think about a specific individual, say Tom Brady, a successful quarterback from the New England Patriots. Despite the focus on the individual, his salient affiliation to the group of US-American Football players may activate the same associated stereotype content and emotions

as thinking about the group does. This should be most likely if the only thing one knows about Tom Brady is this particular group affiliation.

Interestingly, while the SCM posits that groups can be uniquely located in the two-dimensional space, this is likely not true for individuals. This is because individuals may belong to many different groups. To illustrate, consider Tom Brady who may be perceived as belonging to the group of American Football players. Alternatively, individuals might categorize Tom Brady into the group of rich people. Importantly, different group affiliations may result in different emotional reactions: while the category football player may trigger admiration, the category rich may trigger envy. Following this reasoning, which specific emotion is evoked for a specific individual depends on which group association is salient.

Based on individuating information. While we started our review with the SCM, it is important to note that the literature on person perception has independently developed a long history discussing two dimensions that describe how people perceive individuals without a specific focus on group affiliation. These dimensions bear important similarities to warmth and competence. For example, Asch (1946) presented a warm-cold dimension, which he distinguished from more competent related adjectives. The more recent Big Two approach posits that person perception can be captured by the two core dimensions communion (someone's intentions) and agency (someone's ability to act upon these intentions; Abele & Wojciszke, 2007). Both the group and the person perception literature thus allow for the inference that the same emotions evoked in group perception may similarly be evoked when judging individuals.

When meeting others, facial information is often more readily available than explicit information about group affiliation or individuating personality information. Faces immediately capture individuals' attention (Cerf, Frady, & Koch, 2009; Crouzet, Kirchner, & Thorpe, 2010) and are widely used to form a first impression about a person (Bar, Neta, & Linz, 2006; Willis & Todorov, 2006), and group affiliation (Martin & Macrae, 2007). In an

attempt to identify the underlying dimensions of face evaluation, Oosterhof and Todorov (2009) showed that the two dimensions trustworthiness and dominance best explain face evaluations. Of interest, these two dimensions can be similarly characterized as an evaluation of someone's intentions (trustworthiness) and the ability to put these intentions into action (dominance). The trustworthiness-dominance conceptualization thus bears important similarities to the two-dimensional warmth-competence (Fiske et al., 2002) and communion-agency (Abele & Wojciszke, 2007) conceptualizations discussed before. This conceptual overlap allows for the intriguing speculation that faces, too, may elicit the emotions most prominent in group perception in a similar vein and might be remapped onto the two-dimensional space of warmth and competence.

Yet, just as individuals may not be uniquely located in the SCM-two-dimensional space simply because they belong to different groups, faces may similarly afford the elicitation of several emotions simultaneously, as we will discuss next.

Envy, Admiration, and Fear

Arguing from the group perspective, it can be assumed that depending on what group Tom Brady is cognitively affiliated to, for instance, either admiration (Tom is an American sports hero) or envy (Tom is rich) may be evoked. However, when confronted with an unknown individual, a distinction between admiration and envy might be less clear because of lacking an explicit group affiliation that results in the activation of stereotypical knowledge. This notion is consistent with conceptualizations of envy as a two-faced emotion (e.g., Smith & Kim, 2007; van de Ven, Zeelenberg, & Pieters, 2011) that may be fueled by benign or malicious intent. Interestingly, benign envy is conceptualized in close proximity to admiration (Foster et al., 1972; Silver & Sabini, 1978; Smith & Kim, 2007).

The SCM associates high-competence and low-warmth with envy. But if envy and admiration are easily grouped together, it appears desirable to include a fifth emotion in our set of studies that may better capture the low-warmth but high-competence spot. Cuddy and

colleagues (2007) noted that when thinking of groups, a lack of warmth likely elicits the feeling of fear, as fear is elicited by perceived threat (Frijda, Kuipers, & ter Schure, 1989) and thus associated with low warmth. Consistent with this notion, in one of their studies Cuddy and colleagues (2007, Study 4) observed a negative correlation between perceived fear and perceived warmth. Interestingly, the authors also observed a positive correlation between perceived fear and competence on the participant level. Fear might thus fill the spot of low warmth and high competence. Indeed, it appears sensible to assume that individuals experience fear if confronted with a person who does not have their best interest in mind (low warmth) and may act accordingly (high competence). Against this background, we opted to include fear as a fifth emotion in our work.

What Faces Look Like That Evoke an Emotion

A seemingly straight forward attempt to investigate what the face of an individual who evokes a specific emotion looks like could be to present different faces and ask to what degree they evoke specific emotions. However, such an attempt would come with many shortcomings, which would render results difficult to interpret. Using the beforementioned approach entails many confounding variables, such as a person's styling or posing. Furthermore, because of the potential salience of the stereotyped groups in such an experiment, results would rather inform about stereotypes about groups than about (the face of) a person that evokes a specific emotion. It would therefore remain unclear which visual cues are actually driving the results. The solution to getting around these problems is that we put the cart before the horse. Thus, we use a fully data driven method—the reverse correlation method—to extract those facial characteristics that are responsible for evoking specific emotions in perceivers. Instead of establishing a correlation between fixed attributes and participants' responses, reverse correlation methods use the correlation between a fixed response variable and random stimuli. This enables the modeling of the attributes that are causing participants' choice pattern (Todorov, Dotsch, Wigboldus, & Said, 2011).

The upcoming sections will review literature on the reverse correlation technique using the image classification task and how it has been applied to the domain of face perception in order to visualize specific attributes and stereotypes. Moreover, we will discuss the boundaries of the traditional reverse correlation technique and how we implement its basic tenets into a statistical face space while using up-to-date computer graphics.

Extracting Prototypes

Visualization of attributes. People are often able to name specific facial features that they believe belong to a specific category and people often agree on these, such as that a happy face shows an upturned mouth or that men have angular faces. However, faces are more holistic than just the mouth or the curvature and the bigger picture may be hard to name accordingly. Thus, people share a consensus about what prototypical exemplars of specific categories look like (e.g., a happy face) but they lack the ability to articulate what information goes with the category. In an attempt to 'make the ineffable explicit,' Mangini and Biederman (2004) worked with an extension of the reverse correlation image classification technique. They presented their participants with 390 images and in one of the experiment conditions they asked them to indicate whether the image showed a happy or an unhappy face. All images were created using a so-called base face and random noise. The base face was a morph of different male and female faces. For each of the 390 trials, a random noise pattern was created and then applied to the base face, resulting in one slightly distorted version of the base face. After collecting participants' answers, all selected noise patterns for one of the two categories were averaged and the average noise pattern was again applied to the base face, representing the consensus regarding a happy or an unhappy face (see Mangini and Biederman (2004) for the visual results). In the same vein, Kontsevich and Tyler (2004) used the reverse correlation image classification technique to extract what is in the face that leads people to perceive a face as looking happy or sad. In their study they used the portrait of Mona Lisa as the base face because it is one of the best-known examples of an expression at

the ambiguous point between happy and sad. The authors overlaid the original portrait with random noise patterns multiple times and asked participants to rank the emotional expression perceived on the portrait with the four categories sad, slightly sad, slightly happy, or happy. Averaging the choices for all sad and all happy choices resulted in two average noise patterns that were applied to the original portrait of Mona Lisa. Because the resulting prototype consists solely of the information of how the participants classified the noise pattern that was applied on the base image, this technique enables the facial characteristics that they used to derive their decisions to be extracted in a data driven way.

Visualization of stereotypes. Adapting the reverse correlation image classification technique to the domain of stereotyping, Dotsch, Wigboldus, Langner, and van Knippenberg (2008) visualized Dutch participants' internal representation of Moroccans, a highly stigmatized immigrant group in the country of data collection (Netherlands). As their study did not use two different categories (e.g., happy versus unhappy) but only a single category (i.e., the typical Moroccan), Dotsch and colleagues (2008) complemented each random noise pattern with its inversion. A specific pixel appearing dark in the original noise pattern appears brighter in the inverted noise pattern, and vice versa. For each trial, the stimulus material therefore consisted of two identical images of the base face, one of which the original noise pattern was applied to while the inverted noise pattern was applied to the other. Both images were simultaneously presented to participants whose task was to repeatedly indicate which of the two versions was the more Moroccan-looking face. By averaging the noise patterns of all faces that participants had chosen, a classification image was calculated. This technique allowed visualization of the stereotype of what a Moroccan person looks like.

This reverse correlation technique has since been widely used by different research groups for different questions in the domain of social psychology (Brown-Iannuzzi, Dotsch, Cooley, & Payne, 2016; Gunaydin & DeLong, 2015; Imhoff, Woelki, Hanke, & Dotsch, 2013; Kunst, Dovidio, & Dotsch, 2018; Ratner, Dotsch, Wigboldus, van Knippenberg, &

Amodio, 2014; Young, Ratner, & Fazio, 2014) and has become a powerful tool to investigate stereotypes. Compared to openly asking about people's stereotypes, the reverse correlation image classification technique is more implicit and thus less susceptible to social desirability considerations. Moreover, no a-priori assumptions are needed regarding which facial information people associate with certain groups because this information will be extracted a posteriori from the data. Finally, this reverse correlation technique allows us to visualize a mental image that individuals may not be able to consciously access or express, even if they wished to.

In this paper we offer an advancement of the traditional reverse correlation technique that preserves the technique's benefits and original notion, that comes along with new possibilities, especially if the aim of the research is to go beyond the mere visualization of prototypes. Here, we outline some of the boundaries that we aim to exceed with this advanced technique. First, the resulting noise pattern from a traditional reverse correlation study only results in a meaningful prototype if applied to the same base face it has been developed upon. To illustrate, applying the noise pattern extracted from the Moroccan classification task to another face may not result in the perception of a Moroccan-looking face. This is because the extracted noise pattern is the result of the unique combination of characteristics of the specific base face used in the classification task and the noise pattern extracted from this task. A dark pixel that matches the corner of the right eye in the original base face and thus bears diagnostic information because it renders the surrounding of the eye darker might result in a dark spot in the white dermis of the eye when applying the noise pattern on another face. In other words, to get an informative result, a specific noise pattern can be applied only to the face it has been originally derived from. Second, the extracted noise pattern produces a static image, meaning the technique does not allow us to produce multiple variations of faces that look more or less like a prototypical member of a specific group (e.g., more or less like the typical Moroccan). Finally, due to the nature of the noise patterns, the resulting visualization

is a grainy, black and white image, and thus reflects more of an approximation than a true internal representation (e.g., see Dotsch et al., 2008). The resulting visualizations of a reverse correlation image classification task (such as 'the typical Moroccan') may thus not be used as realistic stimuli. Together these limitations forestall the implementation of more complex research designs that wish to generalize not only across participants but also across stimuli (Judd, Westfall, & Kenny, 2012, 2017). To illustrate, if a researcher wants to present his or her participants with multiple faces, such as typical Moroccan looking faces or with atypical Moroccan faces, she/he needs multiple face IDs and the possibility to manipulate each of them in both directions (i.e., looking more Moroccan and less Moroccan) to deconfound face ID and stereotype content.

In order to account for the abovementioned boundaries, we implement the basic tenets of the traditional reverse correlation technique (Dotsch et al., 2008) in a statistical face space (Paysan et al., 2009) while using up-to-date computer graphics (Walker & Vetter, 2009, 2016). The advanced reverse correlation technique enables the extraction of realistic stimuli that researchers can fruitfully rely on to investigate, for instance, person perception in a more ecologically valid fashion and with more complex research designs.

From random noise patterns to random vectors

The face space approach assumes that every face can be represented as a specific point in a multidimensional space. Two faces that are more similar to each other are also closer to each other in this multidimensional space. Conversely, a third face that differs from these two similar faces is located further away in the multidimensional space (Valentine, 1991).

Based on this assumption, Paysan and colleagues (2009) created a statistical face space. For this purpose, they took 3D face scans of 100 female and 100 male volunteers who displayed a neutral facial expression. Each 3D face scan was then mathematically represented by 53490 3D vertices coding for shape and 53490 values coding for the color of these vertices. To extract the dimensions that best explained the variance among all the faces, two

principal component analyses were conducted. This led to a 199-dimensional space for shape and a 199-dimensional space for color (Paysan et al., 2009). Every face is located at a specific position on each of the 199 shape and color dimensions and can be described as a vector pointing from the center of the shape and color face spaces to its actual position in the multidimensional spaces (i.e., individual face). The center of both spaces is expressed by the value of zero on all dimensions (i.e., average face or zero vector face). By randomly combining values for each dimension in the shape space and in the color space, new random faces can be created (i.e., random face). If the same random vector is added once and subtracted once from the average face, two random faces are created that oppose each other in the multidimensional face spaces.

To extract a shared representation of a specific group or a characteristic, we developed a reverse correlation image classification technique with stimulus material that is randomly located in the statistical face space. During the study, participants will be presented with two faces on the same page. One of the faces is the average face with the random vector added (the original noise pattern), and the other is the average face with the random vector subtracted (the inverted noise pattern). Participants are asked to choose which of the two faces is more likely to fit into a specific category. In the next step, the prototype-vector for each participant or for all participants can be calculated and visualized. Of note, the very same vector can be applied to *any* face in the face space, moving the face in the direction of the extracted vector.

The Present Research

In this paper we aim to gain insights into what a face looks like that evokes a specific emotional reaction in perceivers. As a means to an end we here present an advanced reverse correlation technique, combining the traditional reverse correlation's basic tenets with a statistical face space and up-to-date computer graphics that enables visualization of

prototypes in a highly realistic manner, application to multiple photographs of faces and comparison of different prototypes with each other.

In five studies, we test three main hypotheses. Hypothesis 1 holds that our technique to extract emotion prototypes and apply them to real face photographs is successful in that they will trigger the respective emotion in perceivers (Studies 1a, 1b and 2). Following the reasoning that (benign) envy is conceptualized in close proximity to admiration, with Hypothesis 2, we predict that someone who evokes envy will look very similar to someone who evokes admiration (Studies 1 – 4). Hypothesis 3 holds that faces that trigger one of the SCM emotions can be consistently located in the dimensional space spanned by warmth and competence, except for envy. More specifically, we predict that faces that elicit the feeling of admiration will be rated as high on warmth and on competence; faces that elicit the feeling of disgust will be rated low on warmth and on competence; faces that elicit the feeling of pity will be rated high on warmth and low on competence; different from the SCM, we predict that faces evoking envy will be rated high on warmth and high on competence; finally, faces that elicit the feeling of fear will be rated low on warmth but high on competence (Study 5a and 5b). All studies were approved by the university's institutional review board (IRB).

In all our studies, our sampling approach was to strive for a sample as diverse as possible. We therefore collected all data online. Our restrictions were that the participants had to be at least 18 years old and that English was their mother tongue. Thus, the obtained sample is diverse in terms of gender, age and background, although our sample is mostly limited to participants from English-speaking countries due to our language restrictions. The sample is therefore clearly grounded in Western cultures and the results can therefore be generalized in particular to this cultural background.

Studies 1a and 1b – Developing Emotion Prototypes

The aim of Study 1a was to extract the socially shared representations of faces that elicit the emotions of admiration, envy, pity, and disgust. Additionally, we extracted the

socially shared representation of a face that elicits the emotion of fear in Study 1b. In order to achieve this goal, we developed an advanced reverse correlation technique, combining an image classification task (e.g., Dotsch et al., 2008) with a statistical face space (Paysan et al., 2009) and up-to-date computer graphics (Walker & Vetter, 2009, 2016). For each emotion we aimed to extract, we used two different sets of random faces that enabled us to test the reliability of the technique.

Method

Participants and design. Previous research using the reverse correlation paradigm (e.g., Dotsch & Todorov, 2012; Dotsch et al., 2008; Imhoff et al., 2013; Kunst et al., 2018) reveals that 20 to 30 participants are needed per cell to produce consistent results. We based our sampling approach on the upper range of these studies, aiming for 30 participants for each stimulus set and a total of 60 participants for each emotion (two sets each). This resulted in a desired sample of 240 participants in Study 1a (four emotions), and 60 participants in Study 1b (one emotion). In order to reach the desired sample sizes in case of any exclusions, we slightly oversampled in both studies: we recruited 249 participants (142 female, 106 male, 1 no answer; M_{Age} = 37.30, SD_{Age} = 13.45) for Study 1a (four emotions) and 63 participants (24 female, 39 male; M_{Age} = 35.00, SD_{Age} = 13.27) for Study 1b (one emotion).

Study 1a used a 2 (stimulus set: random face pairs 1-400 vs. random face pairs 401-800) by 4 (emotion: admiration vs. envy vs. pity vs. disgust) between-subjects design. In Study 1b we used the same stimulus sets but only one emotion (i.e., fear). Both experiments were set up as online studies and were run on Prolific Academic (www.prolific.ac).

Material. For both stimulus sets we created 400 random vectors in the multidimensional face spaces. Half of these vectors varied on the first 50 shape components, and the other half varied on the first 50 color components. To generate our random vectors, we created four subsets of 100 random vectors each. We ensured that every component of the

vector had similar variance by setting the mean of the random distributions for each component to 0 and the standard deviation to 0.3.

To create the face pairs for the trials, we used a base face that is a morph consisting of the 200 3D scans of participants' faces displaying a neutral facial expression that the Basel Face Model is built upon (Paysan et al., 2009). Each random vector was added as well as subtracted from the base face. Because the used base face can be defined as a vector that has, by default, the value 0 for each component, it is located in the center of the multidimensional face space. By adding and subtracting a random vector once, we move the base face in opposite directions from each other in the multidimensional space. Each random vector thus results in two opposing random faces. In a final step, the modified faces were frontally rendered on a white background. The size of the images is 524 by 524 pixels. See Figure 1 for an illustration of the two spaces where exemplary random vectors have been applied onto the base face.

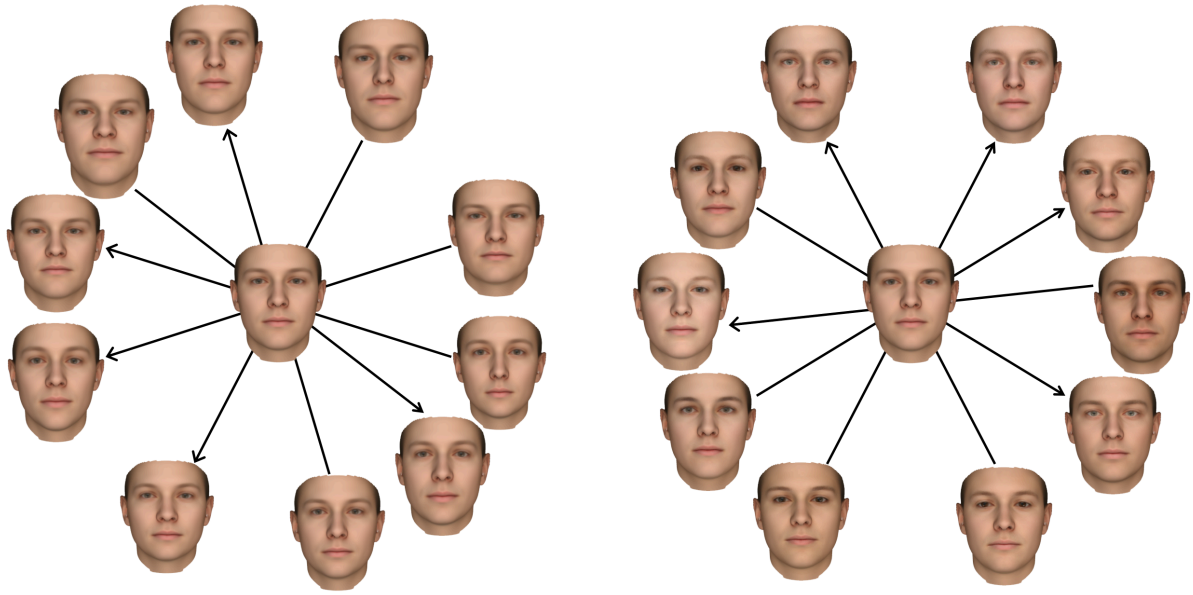


Figure 1. The left (right) shows the multidimensional shape (color) space in a parsimonious way. In the middle of both spaces is the base face. Each arrow symbolizes a specific random vector that is added once (arrowhead) and subtracted once (arrow end) from the base face. The random vectors on the left side manipulate the shape of the base face only (i.e., the shape space) and the random vectors on the right side manipulate the color of the base face only (i.e., the color space).

Procedure. Participants were welcomed, told that the study was dealing with first impressions, and were asked to provide informed consent. Participants were then told that we were interested in facial attributes and their connection to emotions. On the next page participants read that they would subsequently work on a number of trials, each consisting of two faces on the screen, one on the left side and one on the right side. Images were presented horizontally next to each other in the middle of the screen; arrangement (left or right) was random. Depending on the emotion condition, participants were asked to indicate which of the two faces was more likely to evoke the emotion of admiration, envy, pity, disgust, or fear, respectively. Participants provided their answers by clicking on the respective face. Each trial ended with a fixation-cross in the middle of the screen presented for 25 milliseconds, dividing each trial from the following one. For each participant this image classification task consisted of 400 trials in total. In 200 trials face pairs solely differed regarding shape information, and in 200 trials face pairs solely differed regarding color information. Participants started with the shape trials and then worked on the color trials. The order within the shape trials and within the color trials was random and there was a break page after every fifty trials.

At the study's end, we asked participants to indicate what kind of device they had used to complete the study and whether they had encountered any problems during study completion (e.g., the images loaded slowly). Finally, we collected participants' demographic data (i.e., age, gender, English skills, and their education) and gave them the opportunity to comment on the study before they received their payment code.

Results

In order to extract the visual representations of faces that evoke admiration, envy, pity, disgust, and fear, respectively, the following steps were executed. First, we averaged all the random faces[1] selected by a participant resulting in an individual prototype for each

[1] From a mathematical perspective, we averaged the vectors underlying the faces. For the sake of understandability, we refer to vectors as "faces".

participant, representing the participant's internal representation of a face that evokes the respective emotion. Second, we averaged these prototypes across all participants within the same set-by-emotion condition, resulting in two prototypes for each emotion (i.e., two prototypes each for faces that evoke admiration, envy, pity, disgust, and fear, respectively). To investigate whether the two prototypes per emotion result in a similar position within the multidimensional face space, we calculated weighted correlations[2] between the two prototypes. The weighted correlations between stimulus set 1 and stimulus set 2 range from r_{envy} = .82 to r_{fear} = .97. Given these high correlations, we next calculated the average prototype separately for each emotion. This final step resulted in five different prototypes of a face, each evoking a different emotion (admiration, envy, pity, disgust, and fear).

Table 1
Correlation-matrix for the different extracted emotion-vectors from Study 1. In the brackets are the weighted correlation between (first) the shape, and (second) the color components between two indicated prototype-vectors.

	Envy	Pity	Disgust	Fear
Admiration	.97 (.87, .99)	-.25 (-.58, -.70)	-.78 (.05, -.97)	-.22 (.28, -.65)
Envy		-.12 (-.38, -.71)	-.85 (-.06, -.98)	-.33 (.14, -.66)
Pity			-.38 (-.81, .61)	-.85 (-.93, .20)
Disgust				.76 (.95, .77)

To investigate the relationship between the five emotion prototypes, we calculated weighted correlations (see Table 1). Two combinations resulted in positive correlations: the admiration prototype (i.e., a face that evokes the emotion admiration) and the envy prototype, r = .97; and the disgust prototype and the fear prototype, r = .76. Thus, the admiration and the envy prototype on the one hand and the disgust and fear prototype on the other hand are located in a highly similar direction in the multidimensional face space. The upper row of Figure 2 visualizes the prototypes. The lower row visualizes the prototype-vector added to a

[2] Every component was weighted by the amount of variance explained in the Basel Face Model (Paysan et al., 2009).

3D estimate of a male face from the Basel Face Database (BFD; Walker, Schönborn, Greifeneder, & Vetter, 2018) and rendered back to 2D.

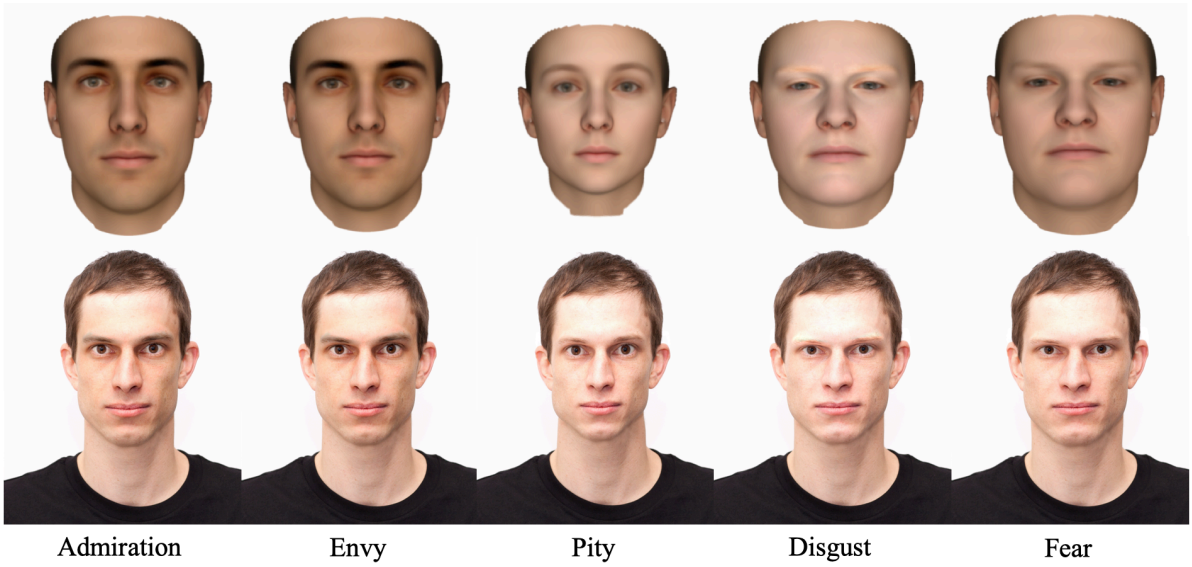


Figure 2. Visualization of the prototypes extracted in Study 1a and 1b. The upper row presents prototype-vectors added to the average face (Blanz & Vetter, 1999) and the lower row the prototype-vectors added to an example ID from the BFD (Walker et al., 2018).

Discussion

Study 1 aimed to extract the socially shared representations of faces that evoke the emotions admiration, envy, pity, disgust, and fear in perceivers. In order to achieve this aim, we used an advanced reverse correlation technique in which participants repeatedly indicated which of two faces was more likely to evoke the respective emotion. By averaging participants' choices, prototypes were extracted that are meant to evoke the respective emotion.

From a methodological perspective, it should be noted that we used two different stimulus sets to extract each of the five different emotion prototypes. Nevertheless, the two resulting prototypes per emotion point in very similar directions in the face space. Thus, although different random faces were used as stimuli in the two sets, the resulting prototypes for the same emotion are highly similar. These findings attest to a high degree of reliability and reproducibility.

Initial informal visual inspections of the prototypes (see Figure 2) revealed that the shape of the *admiration* prototype is elongated, has strong eyebrows and eyes wide open. The overall texture is darker than the other prototypes, except for the envy prototype. The shape of the *envy* prototype is also elongated, has strong eyebrows, eyes wide open, although seemingly to a lesser degree than the admiration prototype, and the overall texture is darker as well, similar to the admiration prototype. Consistent with the reported prototype correlations, the admiration and the envy prototype thus look similar in many respects.

The shape of the *pity* prototype is small and roundish. The eyebrows are more curved than those of the other prototypes. The overall texture of the face is rather pale.

The shape of the *disgust* prototype is wide, the corners of the mouth are pointing downwards, the eyes are more closed, and the eyebrows are thin and straight. The overall texture is also rather pale. The *fear* prototype shares high similarity with the disgust prototype, especially regarding the shape of the face. The texture of the fear prototype, however, is less pale than the disgust prototype. Again, this is consistent with the reported prototype correlations.

In sum, visual inspection indicates that the admiration and envy prototypes look similar but are distinguishable from the other three prototypes. The pity prototype looks rather distinct from all the other emotion prototypes. Finally, the disgust and fear prototypes look similar especially regarding shape information, but are distinguishable from the other prototypes. These results suggest that faces evoking the emotions of admiration and envy are represented similarly, and faces evoking the emotions of disgust and fear are represented similarly. Pity is well distinguishable from the other four prototypes.

These results differ from findings on group research where the emotions admiration, envy, pity, and disgust are suggested to cluster nicely in four separate quadrants (Fiske et al., 2002). However, as we will detail later, there is good reason to assume that while faces in

themselves allow for important inferences, further contextual information is needed to decide whether someone should be admired versus envied, or loathed versus feared.

Study 2 – Validation of Emotion Prototypes

Study 2 sought to validate the extracted prototypes and to investigate to what degree participants distinguish between the different emotion prototypes. To this end, we apply the prototype-vectors to real face portraits in Study 2, and asked participants to indicate to what extent each of the portraits elicits the five emotions admiration, envy, pity, disgust, and fear.

Method

Participants and design. A priori power calculation for a mixed-effects-model (Westfall, 2016) resulted in an aspired sample of 120 participants in order to detect a small to medium effect size of $d = 0.4$ with a power of .80 using 25 different face identities (face IDs). We collected data from 125 participants. We excluded three participants who failed the attention check and one participant who indicated that there was reason not to use their data (e.g., because they clicked through the survey without paying attention). This resulted in a final sample of 121 participants (63 female, 56 male, 2 no specified gender) with a mean age of 32.55 years ($SD = 11.02$). We varied emotions and face identity within participants. Emotions are nested in face IDs and face IDs are nested in participants. The study was set up as an online experiment and was run on Prolific Academic (www.prolific.ac).

Material. Twenty-five male face IDs from an extended version of the BFD (Walker et al., 2018) were used. We first reconstructed the faces in the 2D photographs resulting in 3D estimations of the respective faces (Schönborn, Egger, Morel-Forster, & Vetter, 2017), added the five prototype-vectors and finally rendered them back to 2D. Adding the respective prototype-vectors to each of these 25 faces, resulting in 125 faces in total (see lower row in Figure 2 for one example ID).

Procedure. In the first part of the study participants were welcomed, told that the study was about first impressions from faces, and provided informed consent. Participants

were asked to decide spontaneously how much each of the portraits triggers specific emotions in them. In total, every participant saw all 25 face IDs, of which five each were manipulated to evoke one of the five emotions. Directly underneath the face, participants indicated to what extent the face elicited admiration, envy, pity, disgust, and fear respectively on a 9-point Likert scale (1 = *not at all*, 9 = *extremely*). This procedure was repeated for all 25 faces.

At the study's end, we asked participants in an open response format to indicate the five emotions on which they had been asked to rate the faces. This item served as an attention and data quality check. We further asked participants to indicate what kind of device they had used to complete the study and whether they had encountered any problems during study completion (e.g., the images loaded slowly). Finally, we collected participants' demographic data (i.e., age, gender, English skills, and their education) and gave them the opportunity to comment on the study before they received their payment code.

Results

In a first step we z-standardized the ratings for each emotion separately because we were not interested in absolute differences between the emotions, but in the extent to which a specific emotion was correctly perceived (i.e., in line with our manipulations). We then ran an overall model to test whether the admiration [envy, pity, disgust, fear] prototype evoked higher ratings for admiration [envy, pity, disgust, fear] than for the other four emotions. We specified the standardized rating as the dependent variable, participant and face ID as random effects, and the overall contrast as fixed effect. For all trials where the manipulated and assessed emotion matched, the weight for the contrast was set to +0.8. For all the trials where the manipulated and the assessed emotion did not match, the weight for the contrast was set to -0.2. This overall model supports the hypothesis that the five emotion prototypes evoked the respective emotions more strongly than the remaining four emotions, $t(14979) = 11.14$, $p < .001$, $R^2 = .30$.

To investigate whether the admiration [envy, pity, disgust, fear] prototype elicits this specific emotion to a higher degree compared to the remaining emotions, we conducted separate contrast analyses for each emotion prototype. Thus, for each prototype we assigned a weight of +1 if assessed emotion and prototype matched and weights of -0.25 if assessed emotion and prototype did not match. Participants and face IDs were treated as random effects and the (1; -0.25; -0.25; -0.25; -0.25)-contrast as fixed effect. Providing support for the validity of the extracted emotion-vectors, every prototype led to higher ratings on the matching emotion compared to ratings on the four non-matching emotions. The z-standardized means and standard-deviations together with test results are presented in Table 2.

Table 2
Means and standard-deviation of the z-standardized values for the contrast comparisons between the different emotion prototypes in Study 2.

	Matching emotion *M* (*SD*)	Other four emotions *M* (*SD*)	*t*-value	*p*-value	R^2
Admiration	0.14 (1.11)	-0.10 (0.91)	6.86	< .001	.38
Envy	0.12 (1.15)	-0.06 (0.98)	4.81	< .001	.37
Pity	0.01 (1.03)	-0.08 (0.94)	2.59	.010	.34
Disgust	0.22 (1.12)	0.03 (1.00)	4.90	< .001	.34
Fear	0.26 (1.14)	0.02 (0.99)	6.14	< .001	.34

Additionally, for every emotion prototype we compared the matching emotion with every non-matching emotion via four mixed-effects-models per emotion prototype. Participant and face ID served as random effects. The contrast testing for the effect between the emotion that has been manipulated in the face and each of the four remaining emotions served as our fixed effect. Participants' rating on how much these specific emotions had been evoked due to the face served as the dependent variable. The admiration prototype led to higher ratings of admiration compared to pity, disgust, and fear, but not to higher ratings of admiration compared to envy. The envy prototype led to higher ratings of envy compared to pity, disgust, and fear, but not to higher ratings of envy compared to admiration. The pity prototype led to higher ratings of pity compared to disgust and fear, but not to higher ratings

of pity compared to admiration and envy. The disgust prototype led to higher ratings of disgust compared to admiration and envy, but not to higher ratings of disgust compared to pity and fear. The fear prototype led to higher ratings of fear compared to admiration and envy, but not to higher ratings of fear compared to pity and disgust. Means, standard deviation with the z-standardized values and test results are presented in Table 3.

Table 3
Mean, standard deviation and t-values for single comparison mixed-effects-models from Study 2.

Emotion presented	Emotion assessed				
	Admiration	Envy	Pity	Disgust	Fear
			M (*SD*)		
Admiration	0.14 (1.11)	0.06 (1.06)	-0.13 (0.88)	-0.15 (0.87)	-0.19 (0.80)
Envy	0.16 (1.14)	0.12 (1.15)	-0.14 (0.87)	-0.12 (0.92)	-0.12 (0.92)
Pity	0.04 (1.00)	0.02 (1.01)	0.01 (1.03)	-0.16 (0.88)	-0.23 (0.85)
Disgust	-0.20 (0.78)	-0.11 (0.87)	0.14 (1.11)	0.22 (1.12)	0.29 (1.11)
Fear	-0.14 (0.87)	-0.09 (0.86)	0.11 (1.05)	0.20 (1.11)	0.26 (1.14)
			t-values		
Admiration		2.00	5.79*	6.17*	7.48*
Envy	-0.76		5.73*	5.15*	5.17*
Pity	-0.55	-0.21		4.26*	5.56*
Disgust	9.15*	7.23*	1.82		-1.44
Fear	8.11*	7.33*	2.92	1.13	

Note. Mean and standard deviation steam from the z-standardized values. Significant *t*-values below corrected alpha-level of .0025 are marked with an asterisk.

Discussion

To validate the five emotion prototypes (i.e., admiration, envy, pity, disgust, and fear), we applied the emotion prototypes extracted in Study 1 to real face photographs and asked participants to indicate to what degree each face elicited the five emotions of admiration, envy, pity, disgust, and fear in them. Planned contrast comparisons between the different emotions showed that a specific emotion prototype evokes the respective emotion more strongly than the other emotions. Single comparisons mirror the correlational findings from Study 1: The more similar two prototypes are (both visually and statistically as observed in Study 1), the more these emotion prototypes also evoke the unmanipulated but neighboring

emotion (observed in Study 2). Thus, the admiration prototype not only evokes more admiration but also more envy compared to the remaining three emotions, and vice versa (henceforth referred to as admiration-envy similarity). A similar pattern can be observed for the disgust and fear prototypes. The disgust prototype not only evokes more disgust but also more fear compared to the remaining three emotions, and vice versa (henceforth referred to as disgust-fear similarity). As in Study 1, the pattern is different for pity, as a face evoking pity is distinct from the other emotion prototypes.

In sum, the validation of the prototypes that were assumed to evoke the emotions of admiration, envy, pity, disgust, and fear was successful. However, admiration and envy, as well as disgust and fear, were intra-pair not well distinguishable. One possible explanation is that the model mostly captures valence and cannot methodologically distinguish between admiration and envy on the one hand and between disgust and fear on the other hand. Alternatively, the neighboring locations of admiration and envy, as well as disgust and fear, may meaningfully reflect how these emotions are actually triggered from faces. Study 3 tests these alternative accounts—methodological artifact versus meaningful reflection of actual representations—with a new set of unmanipulated faces. Should similar results be obtained with these unmanipulated stimuli, a methodological artifact is unlikely.

Study 3 – Association of Emotions Evoked by non-Manipulated Faces

The finding that admiration and envy, as well as disgust and fear, are not easily distinguishable calls for further investigation. Study 3 provides a critical test by using non-manipulated faces with neutral facial expressions of an independent database. In particular, we presented participants with non-manipulated male face photographs from the Chicago Face Database (CFD; Ma, Correll, & Wittenbrink, 2015) and asked them to indicate to what degree the presented persons trigger the emotions of admiration, envy, pity, disgust, and fear in them. If a pattern of findings similar to the one observed in Study 2 emerges, we can rule out that the findings are due to our technique or the specific facial material employed.

Method

Participants and design. An a priori power analysis using PANGEA (Westfall, 2016) for mixed-effects-models yielded a required sample size of 152 participants in order to detect a small-to-medium effect size ($d = 0.35$[3]) with a power of .80 when using 25 face stimuli. In total 159 participants finished the study. We excluded three participants who failed the attention check and three participants who indicated that there was reason not to use their data (e.g., because they clicked through the survey without paying attention). This leads to a final sample of 153 participants (100 female, 53 male) with a mean age of 35.33 years (SD = 13.27). The study was set up as an online survey that was distributed through the platform Prolific Academic (www.prolific.ac).

Material. We randomly selected 25 white male faces from the CFD (Ma et al., 2015) that served as stimuli. All images were scaled to a size of 728 by 512 pixel with 240 dpi.

Procedure. We advertised the study as Emotions and Faces and asked participants to spontaneously indicate to what extent each of several portraits elicits certain emotions in them. The procedure was the same as in Study 2, again using a total of 25 faces, which were rated on the five emotions (i.e., admiration, envy, pity, disgust, and fear) using a 9-point Likert scale (1 = *not at all*, 9 = *extremely*).

Results

We first calculated for every face a mean score of how much it triggered each of the five emotions. Second, we calculated the correlations between the different emotions. The face ID served as a unit of analysis because we were interested in the overall connection between different emotions on the level of faces but not on the level of participants. The highest correlations were observed for the combination of admiration and envy ($r = .95$), and for the combination of disgust and fear ($r = .85$). Thus, a face that evokes admiration also

[3] Because we wanted to make sure to detect the smaller effect for pity, we opted for a smaller effect-size in Study 3 compared to the $d = 0.4$ used in Study 2.

evokes envy, and vice versa. Likewise, a face that evokes disgust also evokes fear, and vice versa. The correlation matrix is presented in Table 4.

Table 4
Correlation matrix for emotions in non-manipulated faces in Study 3.

	Envy	Pity	Disgust	Fear
Admiration	.95*	-.36	-.73*	-.57
Envy		-.18	-.56	-.41
Pity			-.66*	-.39
Disgust				.85*

Note. Data structure is in wide format and correlations are calculated with faces as unit of analysis. Significant correlation coefficients below a corrected alpha-level of .0025 are marked with an asterisk.

Predicting the results from Study 2 with the correlation pattern of Study 3. Next, we tested whether the degree to which two emotions are correlated in non-manipulated faces (Study 3) can predict how well people are able to distinguish between the emotions when we manipulate them in faces (Study 2). To test this, we ran a linear regression analysis with the correlation coefficients between different emotions elicited by non-manipulated faces in Study 3 as the predictor variable and the absolute *t*-values from Study 2 as our dependent variable, yielding a highly significant result: $t(18) = -4.24$, $p < .001$, $R^2 = .50$. This indicates that the less [more] likely it is that two emotions are triggered simultaneously by non-manipulated faces (i.e., lower [higher] correlation between emotions in Study 3), the more [less] participants distinguish between the emotion prototypes (i.e., higher [lower] *t*-values in Study 2). Thus, the pattern of how the emotions admiration, envy, pity, disgust, and fear are concurrently evoked in non-manipulated faces is nicely captured in the prototypes we extracted in Study 1 and validated in Study 2.

Discussion

In order to rule out the possibility that the observed findings from Study 2 (i.e., admiration-envy and disgust-fear similarity) was a methodological shortcoming, in Study 3 we presented participants with non-manipulated faces from an independent database. We asked participants to indicate how much the persons in the photographs trigger the emotions

of admiration, envy, pity, disgust, and fear. Consistent with Study 2 we found that faces that evoke the feeling of admiration likely evoke the feeling of envy, and vice versa. Similarly, faces that evoke the feeling of disgust likely evoke the feeling of fear, and vice versa. Pity was again distinguishable from the other four emotions.

Supporting this description, the degree to which two emotions are perceived to co-occur when perceiving non-manipulated faces predicts how similar the emotional reactions towards the respective two prototypes we extracted in Study 1 are. Thus, the degree to which two emotions are evoked by unmanipulated faces is also reflected in the prototypes we extracted in Study 1. This result provides support for the validity of the presented advanced reverse correlation technique, and for the reliability of the observed pattern of how the emotions are concurrently triggered from faces.

Study 3 rules out that the finding that admiration and envy, as well as disgust and fear, are not easily distinguishable from faces is due to our technique or the face stimuli we used. However, because Studies 1 to 3 relied on facial material, it remains an open question whether this finding reflects some peculiarity of face perception. Study 4 addresses this question by investigating whether and how the emotions admiration, envy, pity, disgust, and fear co-occur on a conceptual level.

Study 4 – Conceptual Associations between the five Emotions

The aim of Study 4 was to test whether the admiration-envy, and disgust-fear similarity is specific to face perception or can be observed on a conceptual level, too. Thus, we aimed to investigate how the emotions admiration, envy, pity, disgust, and fear towards another person co-occur with each other on a conceptual level.

Method

Participants and design. We calculated an a priori power analysis using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007) assuming a small to medium effect size (d = 0.35) with a power of .80 and a corrected Alpha-level of

.0016, resulting in a required sample size of 134 participants. In total we collected data from 150 participants. We excluded one participant who failed the attention check and one participant who indicated that there was a reason not to use their data (e.g., because they clicked through the survey without paying attention). This led to a total sample of 148 participants (50 male, 96 female, 2 no answer; M_{Age} = 34.22, SD_{Age} = 12.41). The study was set up as an online experiment and was run on Prolific Academic (www.prolific.ac).

Procedure. At the beginning of the study, participants were welcomed and told that this study investigates how different emotions are related to each other, and were asked to provide informed consent. Next, we told participants that they would be asked to imagine that a stranger induces a certain feeling in them and that we would ask them to indicate to what extent that person might also induce other feelings in them. Furthermore, we told participants that we were interested in first impressions and that there were no right or wrong answers.

The main part of the study consisted of five consecutive pages. The instruction on each of the five pages read "Please imagine feeling admiration [envy, pity, disgust, fear] for a stranger." Below this initial statement the remaining four emotions were presented and participants were asked to indicate to what extent they also experienced each of the other four emotions on a 9-point Likert scale (1 = *not at all*, 9 = *strongly*). The four emotion items per page were presented in alphabetical order. The order of the five pages was random.

Next, we asked participants what kind of device they had used to fill out the questionnaire and to describe the task they were asked to perform in the study in one or two sentence(s), serving as an attention check. Finally, we collected participants' demographic data (i.e., age, gender, English skills, and their education), asked whether there were any problems during the study, how carefully they followed the instructions, and if there was any reason not to use their data. On the last page participants could leave comments regarding the study, were thanked, and redirected to prolific for payment.

Results

This study aimed to test whether the admiration-envy and disgust-fear similarity we consistently found in faces would also be apparent on a conceptual level. Therefore, for each emotion that had been presented to the participants (i.e., presented emotion) we calculated a mean rating of how likely it was that the remaining four emotions were evoked (i.e., assessed emotion). Table 5 shows the mean ratings for all combinations of presented and assessed emotion. The highest likelihood of two emotions co-occurring were reported for admiration and envy. If admiration was presented, it was most likely that envy would also be felt, too (M = 5.23, SD = 2.45). If envy was presented, it was most likely that admiration would also be felt (M = 5.40, SD = 2.38).

If disgust was presented, the likelihood that fear would also be felt was rated rather high compared to the other combinations (M = 3.80, SD = 2.36). If fear was presented, the likelihood that disgust would also be felt was rated rather high compared to the other combinations (M = 3.85, SD = 2.57).

When disgust [fear] was presented, ratings for the likelihood of experiencing pity at the same time were descriptively as high as experiencing fear [disgust]. If disgust was presented, the likelihood that pity would also be felt was rated at a similar level as experiencing fear (M = 4.05, SD = 2.49). If fear was presented, the likelihood that pity would also be felt was rated at a similar level as experiencing disgust (M = 3.86, SD = 2.64).

Table 5

Mean rating for conceptual similarity between different emotions from Study 4.

Emotion presented	Emotion assessed				
	Admiration *M* (*SD*)	Envy *M* (*SD*)	Pity *M* (*SD*)	Disgust *M* (*SD*)	Fear *M* (*SD*)
Admiration	-----	5.23 (2.45)^{a}	1.72 (1.43)^{bc}	1.59 (1.24)^{b}	2.04 (1.63)^{c}
Envy	5.40 (2.38)^{a}	-----	1.90 (1.54)^{b}	2.59 (2.00)^{c}	2.40 (1.85)^{c}
Pity	2.52 (2.02)^{a}	1.55 (1.14)^{b}	-----	2.74 (2.10)^{a}	2.52 (1.94)^{a}
Disgust	1.48 (1.26)^{a}	1.73 (1.45)^{a}	4.05 (2.49)^{b}	-----	3.80 (2.36)^{b}
Fear	2.28 (1.97)^{a}	2.06 (1.67)^{a}	3.86 (2.64)^{b}	3.85 (2.57)^{b}	-----

Note. Different superscripts indicate cell-wise significant differences within rows at corrected alpha-level of .0016.

Additionally, we calculated six paired t-tests within each of the five presented emotions in order to test whether one emotion was more likely to be evoked than another. For example, when participants were asked to imagine someone that evokes admiration, we compared whether it was more likely that this person at the same time evokes more envy [pity, disgust, fear] compared to pity [disgust, fear]. Thus, in total we performed thirty paired-*t*-tests and corrected the alpha-level accordingly to .0016 (see Table 5). In line with results from the face studies, it is most likely that someone who evokes admiration evokes envy compared to pity, disgust, or fear. Someone who evokes envy more likely evokes admiration compared to pity, disgust, or fear. Someone who evokes pity more likely evokes admiration, disgust, or fear compared to envy, although descriptively all ratings are rather low. Someone who evokes disgust more likely evokes pity or fear compared to admiration or envy. Finally, someone who evokes fear more likely evokes pity or disgust compared to admiration or envy.

Predicting results from Study 2 with the mean rating pattern of Study 4. Next, we tested whether the rating about the conceptual similarity of two emotions (Study 4) can predict how well people are able to distinguish between the emotions when we manipulate them in faces (Study 2). To test this, we ran a linear regression analysis with the mean rating of all 20 possible pair combinations from the present study as the predictor variable and the absolute *t*-values from Study 2 as our dependent variable, yielding a highly significant result: $t(18) = -3.61$, $p = .002$, $R^2 = .42$. This indicates that the less [more] similar two emotions are conceptually perceived to be (i.e., lower [higher] ratings in Study 4), the more [less] participants distinguish between the emotion prototypes (i.e., higher [lower] *t*-values in Study 2). Thus, the pattern of how unknown individuals are perceived to trigger two emotions simultaneously is nicely captured in the emotion prototypes we extracted in Study 1 and validated in Study 2.

Discussion

Study 4 tested if the pattern consistently observed with facial material in Studies 1 to 3 replicates on the conceptual level, too. The observed results in Study 4 are consistent with those of the previous studies. Firstly, admiration and envy are perceived as conceptually highly similar. If an unknown person evokes admiration [envy], the likelihood that this person also evokes envy [admiration] was rated as far higher than for any other combination. Secondly, disgust and fear are also perceived as conceptually similar. If an unknown person evokes disgust [fear], the likelihood that this person also evokes fear [disgust] was rated higher than feeling admiration or envy. However, if an unknown person evokes disgust [fear], the likelihood that this person also evokes pity was rated as high as feeling fear [disgust].

Most important, the conceptual rating about how likely it is that two emotions co-occur predicts how well participants are able to distinguish the emotions between the prototypes we extracted in Study 1. This finding is further in line with the assumptions of a dynamic structured social trait space which holds that the conceptual trait space and the perceptual trait space are dependent on each other (Stolier, Hehman, & Freeman, 2018). With this result we achieved strong support for the assumption that the beforementioned pattern of how the emotions admiration, envy, pity, disgust, and fear co-occur with each other is not limited to the domain of face perception but goes beyond and emerges on a conceptual level.

The results for pity did not match perfectly with the results of studies 1-3. On the conceptual level, someone who evokes disgust [fear] is not only more likely to evoke fear [disgust], but also pity. We think that this makes sense given that no contextual and no facial information is available on which participants can rely. Put differently, because faces provide more information than conceptual information only, they also constrain the possible inferences.

Together, Studies 3 and 4 show that the finding that admiration and envy, as well as disgust and fear, are not easily distinguishable is not a methodological artifact from the

advanced reverse correlation technique. We find the same pattern in non-manipulated faces from an independent face database, thus ruling out the possibility that the findings are due to our technique or to the face stimuli we used (Study 3). Furthermore, we find the same pattern when asking about the co-occurrence of admiration, envy, pity, disgust, and fear on a conceptual level, thus ruling out the possibility that the findings are limited to the domain of face perception (Study 4).

Study 5a and 5b – Locating the Prototypes in the SCM

Study 1 extracted facial prototypes that are important in group perception, and Studies 2 to 4 established reliability and validity. Studies 5a and 5b sought to locate the prototypes in the two-dimensional space spanned by warmth and competence in the SCM framework.

Previous research has shown a link between the content of stereotypes and the emotional responses towards the respective groups. It is in the nature of the SCM to focus on this relationship unidirectionally, that is, from content to emotion (Cuddy et al., 2007). Here we focus on the reverse direction, asking, for example, whether a person eliciting the emotion of pity is also characterized as warm but not competent. We address this question by presenting participants with faces that are manipulated to evoke admiration, envy, pity, disgust, and fear and to rate them on the two dimensions of the SCM, warmth and competence. The study was set up as an online survey that was distributed through the platform Prolific Academic (www.prolific.ac). The study was preregistered on AsPredicted.org, see http://aspredicted.org/blind.php?x=yr5h8h

In total, we formulated six partly contradicting hypotheses regarding how the faces manipulated to evoke the beforementioned emotions would be rated on warmth and competence. Thus, we specified six contrasts, which are presented in Table 6.

Table 6
Predicted contrast codes for Study 5 and resulting t-values from Study 5a.

Hypothesis	Dependent trait	Emotion manipulated					*t*-value
		Admiration	Envy	Pity	Disgust	Fear	
I a 1	Warmth	1	- 1	1	- 1	0	11.45
I a 2	Warmth	1	1	1	- 3	0	18.52
I b	Competence	1	1	- 1	- 1	0	8.84
II a	Warmth	1	1	1	- 1.5	- 1.5	4.16
II b 1	Competence	1	1	- 1.5	- 1.5	1	23.17
II b 2	Competence	1.5	1.5	- 1	- 1	- 1	12.09

Note. Degrees of freedom all range between 2400.98 and 2401.86.

In line with the assumptions from the SCM framework, Hypothesis Ia1) predicts that manipulations of admiration and pity lead to higher warmth ratings compared to manipulations of envy and disgust. Hypothesis Ia2) predicts that manipulations of admiration, envy, and pity lead to higher warmth ratings compared to manipulations of disgust. This is in line with the assumptions from the SCM framework except for envy, which we predict to be in the same quadrant as admiration. In line with the assumptions from the SCM, Hypothesis Ib) predicts that manipulations of admiration and envy lead to higher competence ratings compared to manipulations of pity and disgust.

The remaining three hypotheses incorporate fear as a fifth emotion and test whether the emotion fear fits better in the low-warmth-high-competence-quadrant than envy if not groups but individuals are the object of investigation. Specifically, hypothesis IIa) predicts that manipulations of admiration, envy, and pity lead to higher warmth ratings compared to manipulations of disgust and fear. Hypothesis IIb1) predicts that manipulations of admiration, envy, and fear lead to higher competence ratings compared to manipulations of pity and disgust. Hypothesis IIb) predicts that manipulations of admiration and envy lead to higher competence ratings compared to manipulations of pity, disgust, and fear.

Method

Participants and design. Study 5a used a within-subjects design. An a priori power analysis using PANGEA (Westfall, 2016) for mixed-effects-models yielded a required sample size of 100 participants in order to detect a small-to-medium effect size ($d = 0.35$) with a

power of .80 using 25 face stimuli. In total 107 participants took part in the study. We excluded five participants who failed the attention check and one participant who indicated that there was reason not to use their data. This led to a final sample of 101 participants (48 female, 52 male, 1 no specified gender) with a mean age of 32.79 years (SD = 12.07). The study was set up as an online experiment and was run on Prolific Academic (www.prolific.ac).

Study 5b used a between-subjects design (competence or warmth condition). In total, 198 participants completed the study (105 female, 90 male, 3 no specified gender M_{age} = 32.24, SD_{age} = 11.58).

Material. In both studies, we used the same stimulus material as in Study 2. Thus, in total we used 125 stimuli resulting from the 25 male faces from the extended Basel Face Database (Walker et al., 2018), each of which had been manipulated to evoke admiration, envy, pity, disgust, and fear.

Procedure. After providing informed consent, participants read that they would be presented with 25 portraits and asked some questions. The overall procedure was similar to Study 2. Each participant was presented with 25 faces, of which five each were manipulated to elicit one specific emotion.

In Study 5a (within design) below each portrait, participants indicated to what extent the pictured person appeared to be competent, capable (i.e., competence dimension), warm and friendly (i.e., warmth dimension) on a 9-point Likert scale (1 = not at all, 9 = extremely). We counterbalanced the presentation order for the two warmth and the two competence items.

In Study 5b (between design) below each portrait, participants either indicated to what extent the pictured person appeared to be competent and capable (i.e., competence condition) or to be warm and friendly (i.e., warmth condition).

Following the 25 portraits, we asked participants to recall on which traits they had rated the presented images (i.e., attention check). Next, we asked which device participants

used to complete the study and whether they had encountered any technical issues with the images during the study. On the following page we collected demographic data (i.e., age, gender, English skills, and education). Then we asked participants how closely they followed the instructions in the study and whether there was any reason not to use their data. Finally, we gave participants the opportunity to leave any comments regarding the study, thanked them and provided them with the code in order to receive their reimbursement.

Results

The here reported results refer to Study 5a (within design). Results from Study 5b are very similar to those of Study 5a and the summarized results from both studies can be found in Table 7.

In a first step we calculated a combined warmth [competence] index for every participant for every face, r_{min} (99) = .79, $p_{max} < .001$. Mean ratings on the combined warmth and the combined competence indices for each emotion prototype (across all 25 faces) are presented in the first two rows of Table 7.

Table 7
Mean rating of warmth and competence in Study 5 for the within and the between design.

	Admiration *M* (*SD*)	Envy *M* (*SD*)	Pity *M* (*SD*)	Disgust *M* (*SD*)	Fear *M* (*SD*)
Within design					
Warmth rating	4.57 (1.80)^{a}	4.34 (1.71)^{ab}	4.28 (1.79)^{b}	3.13 (1.64)^{c}	3.27 (1.78)^{c}
Competence rating	5.03 (1.75)^{a}	4.97 (1.77)^{a}	4.78 (1.81)^{a}	4.13 (1.80)^{b}	4.13 (1.86)^{b}
Between design					
Warmth rating	4.38 (1.62)^{a}	4.29 (1.63)^{ab}	4.12 (1.63)^{b}	3.14 (1.58)^{c}	2.91 (1.54)^{d}
Competence rating	5.30 (1.66)^{a}	5.23 (1.60)^{a}	5.04 (1.63)^{b}	4.30 (1.67)^{c}	4.36 (1.67)^{c}

Note. Different superscripts indicate cell-wise significance difference within rows at corrected alpha-level of .005.

To test the preregistered hypotheses, we calculated mixed-effects-models. We specified the combined warmth [competence] rating as the dependent variable, participant and face ID as random effects, and the planned contrast as fixed effect. Supporting Hypothesis Ia1) and in line with the assumptions of the SCM, the results reveal that faces that have been manipulated to evoke admiration and pity are rated higher on warmth (M = 4.42, SD = 1.80)

than faces that have been manipulated to evoke envy and disgust (M = 3.74, SD = 1.78); t(2401.19) = 11.45, p < .001, R^2 = .49. However, results also support Hypothesis Ia2), meaning that faces that have been manipulated to evoke admiration, envy, and pity are rated higher on warmth (M = 4.40, SD = 1.77) than faces that have been manipulated to evoke disgust (M = 3.13, SD = 1.64); t(2401.86) = 18.52, p < .001, R^2 = .53. A model comparison between Ia1) and Ia2) shows a better fit of the data with model Ia2), indicating that the envy prototype is perceived as warm, X^2 = 193.13, p < .001.

Supporting Hypothesis Ib) and in line with the assumptions of the SCM, the results reveal that faces that have been manipulated to evoke admiration and envy are rated higher on competence (M = 5.00, SD = 1.76) than faces that have been manipulated to evoke pity and disgust (M = 4.46, SD = 1.83); t(2401.24) = 8.84, p < .001, R^2 = .53.

Supporting Hypothesis IIa) results reveal that faces that have been manipulated to evoke admiration, envy, and pity are rated higher on warmth (M = 4.40, SD = 1.77) than faces that have been manipulated to evoke disgust and fear (M = 3.20, SD = 1.71); t(2400.98) = 4.16, p < .001, R^2 = .52. Supporting Hypothesis IIb1) results reveal that faces that have been manipulated to evoke admiration, envy, and fear are rated higher on competence (M = 4.72, SD = 1.84) than faces that have been manipulated to evoke pity and disgust (M = 4.46, SD = 1.83); t(2401.46) = 23.17, p < .001, R^2 = .56. However, results also support Hypothesis IIb2) meaning that faces that have been manipulated to evoke admiration and envy are rated higher on competence (M = 5.00, SD = 1.76) than faces that have been manipulated to evoke pity, disgust, and fear (M = 4.35, SD = 1.85); t(2401.61) = 12.09, p < .001, R^2 = .54. A model comparison between IIb1) and IIb2) shows a better fit of the data with model IIb2), indicating that the fear prototype is not perceived as competent, X^2 = 124.81, p < .001.

Overall we observed a high correlation between warmth and competence ratings, r = .63, t(2523) = 40.54, p < .001. A person that is rated high on competence is also rated high on warmth, irrespective of the emotion that has been manipulated, r_{min} = .58.

Discussion

Study 5a and 5b investigated how faces that evoke the emotions considered important in group perception (i.e., admiration, envy, pity, and disgust) are rated on the warmth and competence dimension. Additionally, we incorporated fear as a fifth emotion that was assumed to be rated high on competence but low on warmth.

Our studies therefore provide insight into how emotional reactions towards known groups differ from emotional reactions towards unknown individuals. Perhaps most prominently, emotions evoked from faces seem to align on one valence dimension instead of being mixed as the SCM would predict for pity, envy, and, as we hypothesized also for fear. This is to say that high (low) competence ratings go hand in hand with high (low) warmth ratings when judging emotion evoking faces.

Consequently, our findings are only partially in line with the SCM. In line with group level findings, faces that evoke admiration and pity are rated higher on warmth compared to faces that evoke envy and disgust. Faces that evoke admiration and envy are rated higher on competence compared to faces that evoke pity and disgust. However, contrary to SCM but in line with our hypothesis, faces that evoke envy are rated as high on warmth as faces that evoke admiration and pity. This finding is also in line with the coherent results from Studies 1 to 4. A person who is perceived to evoke admiration likely also evokes the neighboring emotion envy, and vice versa. It is therefore also straight forward that a person who evokes admiration and a person who evokes envy are rated similarly regarding their personality.

The results do not support our hypothesis regarding the emotion fear. Our a priori assumption was that faces that are perceived to evoke fear will be rated high on competence but low on warmth. However, results are more in line with the coherent findings of studies 1 to 4. A person who is perceived to evoke disgust is likely to evoke the neighboring emotion fear, and vice versa.

This pattern was found when asking participants to rate the faces on both warmth and competence dimensions simultaneously (Study 5a), and when asking participants to rate the faces only on one of the two dimensions (Study 5b).

General Discussion

This paper investigates what faces that evoke specific emotions look like. Due to the conceptual similarity of the SCM (Fiske et al., 2002) with concepts on person (Abele & Wojciszke, 2007) and face perception (Oosterhof & Todorov, 2009), we used the SCM as a frame of reference by investigating the emotions identified as important in group perception; admiration, envy, disgust, and pity. Additionally, we included fear as an emotion of interest in our studies. Moreover, we here present an advanced reverse correlation technique, combining the image classification task (Dotsch et al., 2008) with a statistical face space (Paysan et al., 2009) and up-to-date computer graphics (Walker & Vetter, 2016). This technique enables us to visualize prototypes in a highly realistic manner in multiple faces and to compare different prototypes with each other.

In Studies 1a and 1b, participants performed an image classification task in which they were repeatedly presented with two faces side by side and were asked to choose the face that evokes the emotion of admiration [envy, pity, disgust, fear] more strongly in them. Based on participants' responses, we extracted prototypes that reflect individuals' representations of faces that evoke admiration, envy, pity, disgust, and fear, respectively. Two sets of different stimuli per emotion yielded very similar results, thus strongly attesting to the technique's reliability. Visual inspection of the prototypes as well as correlational patterns indicate that the admiration and envy prototypes are highly similar (i.e., admiration-envy similarity). Likewise, the disgust and fear prototypes are similar, but to a lesser extent (i.e., disgust-fear similarity).

With the goal of gauging the extracted emotion prototypes' validity, they were applied to a set of real photographs in Study 2. Participants' task was to indicate to what degree each

face evoked the five emotions admiration, envy, pity, disgust, and fear in them. The results provide strong support that each of the extracted prototypes accurately captures the emotion it is meant to reflect (e.g., faces manipulated to evoke admiration were perceived as more admirable than all other emotions combined). Moreover, the data of Study 2 dovetails with the visual inspection of the prototypes and correlational data observed in Studies 1a and 1b. That is, faces manipulated to evoke admiration not only evoked admiration but also envy, and vice versa. Likewise, faces manipulated to evoke disgust not only evoked disgust but also fear, and vice versa.

Given the similarities between envy and admiration, as well between disgust and fear, Studies 3 and 4 sought to investigate to what extent the findings are method- and/or domain-specific. Study 3 relied on a set of non-manipulated portraits from an independent database. As in Study 2, participants indicated to what extent these faces evoked the feelings of admiration, envy, pity, disgust, and fear in them. Results closely resemble those of Study 2, and predict how well participants are able to distinguish between our prototypes. Because the results of Study 2 were replicable with a different, non-manipulated set of faces, Study 3 allows for the conclusion that the pattern observed in Studies 1 and 2 is not an artifact of the specific way we extract facial prototypes but that it is inherent in how emotions are perceived from faces.

A domain-specific uncertainty was whether the observed similarities are specific to the perception of faces. Study 4 therefore went beyond face perception and tested the pattern of emotions on a conceptual level. Participants were asked to imagine feeling admiration [envy, pity, disgust, fear] towards a stranger and to indicate to what degree the other emotions would be evoked concurrently. Results again mimic the pattern from the previous studies, and results of how emotions co-occur on a conceptual level predict how well participants are able to distinguish between our prototypes.

Due to this converging evidence, we conclude that people do have similar representations of someone who evokes admiration and of someone who evokes envy, at least if no additional information is provided. This similarity is further apparent, although to a lesser degree, for the representations of someone that evokes disgust and someone that evokes fear.

Study 5 was conducted to map the emotion prototypes onto the 2-dimensional warmth-competence grid. Participants were presented with the same stimulus material as in Study 2 (i.e., real photographs manipulated to evoke specific emotions) and asked to indicate to what degree they perceived the individuals as competent and (Study 5a) /or (Study 5b) as warm. Results suggest that faces that evoked admiration, pity, or envy were rated higher on warmth and on competence than faces that evoked disgust or fear. Interestingly, we didn't observe "mixed" ratings of warmth and competence. Rather there seemed to be one valence dimension meaning that whoever was judged to be competent, for instance, was also judged to be warm.

In sum, the here presented studies allow for the conclusion that the newly developed advanced reverse correlation technique reliably captures the facial characteristics that evoke specific emotions in perceivers. Moreover, our results suggest an admiration-envy and disgust-fear similarity on the level of individuals. A person who is admired is represented highly similarly to someone who is envied. And a person who is loathed is represented highly similarly to someone who is feared.

Emotions Evoked from Groups versus Faces

In this project we focused on the emotions admiration, envy, pity, and disgust because these are the emotions that, according to the SCM (Cuddy et al., 2007; Fiske et al., 2002), are assumed to be evoked by specific social groups. The original contributions located the emotions in four quadrants by asking for emotional reactions towards specific social groups. The here presented pattern of the admiration-envy similarity appears to be in contradiction

with the SCM-localizations. However, we believe that this contradiction is not genuine. In the context of the SCM, participants are provided with group labels, one by one. A group label provides heuristic stereotypical knowledge that is easily accessible for most individuals in a given culture. Given its stereotypical nature, this information may afford rather non-ambiguous trait ascriptions and emotional reactions (Bodenhausen & Wyer, 1985; Bodenhausen, 1990; Macrae, Milne, & Bodenhausen, 1994). This should be particularly the case if groups are considered one-by-one (as in the SCM), leaving aside that individuals usually pertain to multiple groups at a time. For instance, one specific person may belong to several groups at the same time: poor, White, female, physically disabled, and lawyer. Considered separately, these groups likely allow for non-ambiguous trait ascriptions and emotional reactions; considered in unison, the categorization may be more ambiguous. This is what we believe to be the reason for why the SCM-localizations cannot be exactly reproduced with faces.

Admiration-envy similarity. Across all studies, we found strong evidence that the representation of someone admired and the representation of someone envied are highly similar. This suggests that it may well be the same person or the same face that we admire or envy, depending on the person's deeds, or depending on the perceiver's own motivation. Literature on envy also points out that envy can be two-faced: benign and malicious (Smith & Kim, 2007). Whereas benign envy is conceptualized more closely to admiration, malicious envy is conceptualized more closely to what can be understood as envy in the traditional sense (Smith & Kim, 2007; Van de Ven et al., 2011). Consistent with this conceptual differentiation, some languages explicitly distinguish between benign and malicious envy (e.g., Dutch: *benijden* and *afgunst*). Against the background of the here presented finding that the representation of someone admired and the representation of someone envied are highly similar, we offer the insight that when US-participants think of someone envied, what they have in mind is more in line with benign envy (i.e., conceptualized closer to admiration) than

with malicious envy. A study with a sample where spoken language makes a clear distinction between benign and malicious envy could inform on this matter in more detail.

Disgust-fear similarity. We further observed a disgust-fear similarity, although to a lesser extent. The emotion fear is not part of the original SCM-framework (Fiske et al., 2002). However, because we hypothesized that envy may not be located in the low-warmth, high-competence spot, but would be in close proximity to admiration, we incorporated fear as a fifth emotion in our studies: people who are feared have non-benign intent (low warmth) and can act upon this intent (high competence). We therefore hypothesized that the face of a person who is willing (i.e., low warmth) and able to harm (i.e., high competence) will be better captured by the emotion fear than by envy. Consistent with this hypothesis, the prototype of someone that evokes fear was rated as lower on warmth than the admiration, envy, and pity prototypes (Study 5). However, inconsistent with our hypothesis, the fear prototype was not rated higher on competence than the admiration, envy, and pity prototypes. Throughout four studies our data suggest that someone who evokes disgust and someone who evokes fear are represented similarly, see correlational evidence between the prototypes (Study 1), visual (Studies 2-3), and conceptual testing (Study 4).

It is noteworthy that the here documented disgust-fear similarity is seemingly at odds with literature that discusses fear and disgust as well distinguishable emotions. For example, the two emotions activate different brain areas (Calder, Lawrence, & Young, 2001) and they are linked to different action tendencies (Susskind et al., 2008). Yet, we speculate that this apparent contradiction will dissolve if context is added to the facial stimuli. Without context, a distinction between someone feared and someone loathed is difficult.

Similarity patterns. In some way, the documented admiration-envy similarity and the disgust-fear similarity suggest that emotional reactions based on unknown individuals are rather little differentiated. Yet, they accurately capture whether someone may be approached (those admired, envied, or pitied) or should be avoided (those disgusted or feared). Because

first impressions from faces are formed within milliseconds (Bar et al., 2006), we argue that emotional reactions based on faces are used to gauge another person's intent and capability, and may subsequently be complemented by further contextual information to yield more precise emotional reactions. Moreover, results from Study 5 show a high correlation between perceived warmth and perceived competence for all emotion prototypes. Irrespective of the manipulated emotion, if someone is perceived as high on warmth this person is perceived as high on competence as well, speaking for a halo effect (e.g., Nisbett & Wilson, 1977). On a speculative note, this finding may be interpreted as suggesting that the emotional reaction towards an unknown person is best described by one valence dimension, and that only with more context information can a more nuanced pattern, as in the SCM framework, arise. While in SCM studies the stimuli are rather artificial in their simplicity, they differ from faces as stimuli especially because new faces need to be spontaneously and quickly tested for trustworthiness and dominance. That is because one tries to assess whether one should avoid the person (i.e., high dominance, low trustworthiness) or can approach the person (i.e., low dominance, high trustworthiness). This leads to the conclusion that the first impression of a stranger may therefore look less differentiated than that of a group people have stereotypical knowledge about. It might be that mixed judgements (mixed stereotypes) require a certain amount of cognitive effort and motivation. Further, emotions such as envy might actually precede the judgment. For instance, someone who envies another person for their power or money might feel better about themselves when they vilify this person by assuming a lack of compassion (the rich person as competent but cold).

Using the Best of two Worlds

This contribution presents an advanced reverse correlation technique that combines the image classification task with a statistical face space approach and uses up-to-date computer graphics. In what follows, we further discuss what our new technique adds to the traditional reverse correlation technique (which enables the extraction of visual representations of

stereotypes in an intuitive manner) and to other statistical face modeling techniques (which enable personality traits to be extracted with more realistic results).

Traditional reverse correlation image classification technique. The traditional reverse correlation image classification technique (Kontsevich & Tyler, 2004; Mangini & Biederman, 2004) enables the extraction of internal representations of facial attributes in a fully data-driven way. In order to extract facial stereotypes, Dotsch and colleagues (2008) slightly altered the initial task of classifying one stimulus at a time into different categories to a forced-choice task where one out of two stimuli needs to be classified into a specific category. The technique works as follows. A base face is overlaid with a random noise pattern once and with the inversion of this random noise pattern once. This procedure is repeated multiple times and each of the resulting face pairs is used as an individual trial in the study. Participants' task is to choose which of the two faces is more in line with their internal representation of what the researcher is asking for (i.e., image classification task). By averaging participants' choices (i.e., reverse correlating the answers), a so-called classification image can be created that represents participants' internal representation.

What renders the reverse correlation image classification technique very useful to extract stereotypes is its intuitive nature. Participants are asked to solve a simple forced-choice task. While using random noise patterns, the differences between the two options are unsystematic and ineffable and therefore not prone to social desirability concerns. This enables the visualization of otherwise covert information.

The here presented technique capitalizes on the traditional reverse correlation's intuitive nature that has the distinct advantage of making something visible that would remain covert if the answer was given deliberately. A positive side effect of this is that socially desirable responses are unlikely. What we are omitting from the traditional reverse correlation technique is the use of static random noise patterns. Instead we used random vectors to

modify the base face, combining the benefits of the image classification task with the benefits of using a statistical face space as discussed below.

Statistical face space. Using statistical face modeling techniques, the perception of various personality traits have successfully been modeled in faces. Oosterhof and Todorov (2009), for example, modeled the basic dimensions of face perception, namely trustworthiness and dominance in computer generated faces. Walker and Vetter (2016) further provided evidence that the Big Five personality traits and the Big Two personality traits can be modeled in a new set of real photographs. In these approaches, either randomly created faces (Oosterhof & Todorov, 2009) or 3D scans of real faces (Walker & Vetter, 2016) are presented to participants who are asked to indicate to what degree the presented person looks, for example, trustworthy. With a reverse engineering approach, the direction in the face space that best represents the personality dimension in question can be determined.

These techniques have proven very useful in extracting specific personality dimensions in faces and provides at least three main advantages. First, the technique enables the application of the extracted facial information either to any computer-generated face (Todorov, Dotsch, Porter, Oosterhof, & Falvello, 2013) or to any real face photograph (Walker & Vetter, 2016). Second, operating within a statistical face space results in the visualization of faces without visible artifacts. Third, the extracted vectors can be related to each other.

The here presented technique capitalizes on these three advantages. What we are omitting is the less intuitive location of a face on a Likert scale.

A unified technique. The here presented technique uses the image classification task from the traditional reverse correlation technique and a statistical face space to model the prototypes. The benefit of combining the image classification task with the face space is the task's intuitive nature and that the whole stimulus material as well as the resulting prototypes are vectors that can be visualized as such in the multidimensional face space without visible

artifacts (see upper row of Figure 2). Additionally, by using up-to-date computer graphics (Walker & Vetter, 2016) the resulting prototype-vectors can be applied to any photograph of a face that can be located in the underlying multidimensional face space and rendered back to 2D (see lower row of Figure 2). Moreover, beyond the visualization of internal representations, the technique further affords comparisons between representations. In particular, due to the use of a statistical face space, different vectors can be related to each other (Walker & Keller, 2019). This may provide an idea about how similar or dissimilar prototypes of, for example, personality dimensions (Stolier, Hehman, Keller, Walker, & Freeman, 2018), emotions, or members of specific groups are. The more we know about the face space, the more opportunities arise to locate specific prototypes in the multidimensional space in relation to each other.

Limitations. The present methodological technique is limited in that only facial information that is inherent in the used face space can contribute to the visualization of a prototype. To illustrate, the here used face space consists mainly of information derived from a rather young and predominately White sample. Information not inherent in these faces cannot be used for visualization. The sample thus constrains the possible shape and texture variation. As a result, although we have not yet tested this empirically, the extraction of a prototype of an ethnicity other than White might be difficult. Operating with other, more diverse face spaces could remedy this challenge.

Another limitation concerns the generalizability of the results. We relied on only male target images in Studies 2, 3, and 5. In a future step it could be interesting to investigate whether the extracted prototypes yield similar results when applied to female faces. While speculative, we have reason to believe this as rather likely because a) the stimulus material we used to extract the prototypes was gender-neutral (i.e., the base face the random vectors were added and subtracted from was a morph between 100 male and 100 female faces), and b) the pattern we found for faces as stimulus material has also been found for the mere concepts as

stimulus material, although we cannot state that participants' default representation of a person evoking a specific emotion was not male.

Conclusion

In a time where judgments based solely on appearance have become increasingly prevalent, the aim of this paper was to investigate and to visualize what faces that evoke specific emotional reactions look like, presenting an advanced reverse correlation technique that combines the image classification task with a statistical face space and up-to-date computer graphics. The findings suggest that there are specific patterns of how the emotions admiration, envy, pity, disgust, and fear are interrelated. We consistently found an admiration-envy, and disgust-fear similarity, reflecting that an envied person may look very similar to an admired person, and likewise a loathed person may look very similar to a feared person. The here presented reverse correlation technique reliably captured the facial characteristics that evoke specific emotions in perceivers. The image classification task is very intuitive for participants, which enables visualization of otherwise hidden characteristics with a technique that is less prone to social desirability. The incorporation into a statistical face space and using up-to-date computer graphics further enables the extraction of realistic looking prototypes that can be applied to any face and the prototypes can further be related with each other in multiple ways.

References

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, *93*(5), 751–763. http://doi.org/10.1037/0022-3514.93.5.751

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, *41*(3), 258–290. http://doi.org/10.1037/h0055756

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*(2), 269–278. http://doi.org/10.1037/1528-3542.6.2.269

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (pp. 187–194). USA. http://doi.org/10.1145/311535.311556

Bodenhausen, G. V., & Wyer, R. S. (1985). Effects of stereotypes on decision making and information-processing strategies. *Journal of Personality and Social Psychology*, *48*(2), 267–282. http://doi.org/10.1037/0022-3514.48.2.267

Bodenhausen, G. V. (1990). Stereotypes as judgmental heuristics: Evidence of circadian variations in discrimination, *1*(5), 319–322. Retrieved from http://journals.sagepub.com/doi/pdf/10.1111/j.1467-9280.1990.tb00226.x

Brown-Iannuzzi, J. L., Dotsch, R., Cooley, E., & Payne, B. K. (2016). The relationship between mental representations of welfare recipients and attitudes toward welfare. *Psychological Science*, *28*(1), 92–103. http://doi.org/10.1177/0956797616674999

Calder, A. J., Lawrence, A. D., & Young, A. W. (2001). Neuropsychology of fear and loathing. *Nature Reviews Neuroscience*, *2*(5), 352–363. http://doi.org/10.1038/35072584

Caprariello, P. A., Cuddy, A. J. C., & Fiske, S. T. (2009). Social structure shapes cultural stereotypes and emotions: A causal test of the stereotype content model. *Group Processes & Intergroup Relations*, *12*(2), 147–155. http://doi.org/10.1177/1368430208101053

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, *9*(12), 1–15. http://doi.org/10.1167/9.12.10

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision*, *10*(4), 1–17. http://doi.org/10.1167/10.4.16

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, *92*(4), 631–648. http://doi.org/10.1037/0022-3514.92.4.631

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychology and Personality Science*, *3*(5), 562–571. http://doi.org/10.1177/1948550611430272

Dotsch, R., Wigboldus, D. H. J., Langner, O., & van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, *19*(10), 978–980. http://doi.org/10.1111/j.1467-9280.2008.02186.x

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. http://doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. http://doi.org/10.3758/BF03193146

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 1–7. http://doi.org/10.1016/j.tics.2006.11.005

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902.

http://doi.org/10.1037/0022-3514.82.6.878

Foster, G. M., Apthorpe, R. J., Bernard, H. R., Bock, B., Brown, J. K., Cappannari, S. C., … Foster, G. M. (1972). The anatomy of envy : A study in symbolic behavior. *Current Anthropology*, *13*(2), 165–202.

Frijda, N. H., Kuipers, P., & ter Schure, E. (1989). Relations between emotion, appraisal and emotional action readiness. *Journal of Personality and Social Psychology*, *57*(2), 212–228.

Gunaydin, G., & DeLong, J. E. (2015). Reverse correlating love: Highly passionate women idealize their partner's facial appearance. *PLoS ONE*, *10*(3), 1–10. http://doi.org/10.1371/journal.pone.0121094

Imhoff, R., Woelki, J., Hanke, S., & Dotsch, R. (2013). Warmth and competence in your face! Visual encoding of stereotype content. *Frontiers in Psychology*, *4*(386), 1–8. http://doi.org/10.3389/fpsyg.2013.00386

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. http://doi.org/10.1037/a0028347

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*, 601–625. http://doi.org/10.1146/annurev-psych-122414-033702

Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology*, *110*(5), 675–709. http://doi.org/10.1037/pspa0000046

Kontsevich, L. L., & Tyler, C. W. (2004). What makes Mona Lisa smile ? *Vision Research*, *44*, 1493–1498. http://doi.org/10.1016/j.visres.2003.11.027

Kunst, J. R., Dovidio, J. F., & Dotsch, R. (2018). White look-alikes: Mainstream culture adoption makes immigrants "look" phenotypically white. *Personality and Social Psychology Bulletin*, *44*(2), 265–282. http://doi.org/10.1177/0146167217739279

Lench, H. C., Flores, S. A., & Bench, S. W. (2011). Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: A meta-analysis of experimental emotion elicitations. *Psychological Bulletin*, *137*(5), 834–855. http://doi.org/10.1037/a0024244

Lerner, J. S., & Keltner, D. (2000). Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition and Emotion*, *14*(4), 473–493. http://doi.org/10.1080/026999300402763

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. http://doi.org/10.3758/s13428-014-0532-5

Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology*, *66*(1), 37–47. http://doi.org/10.1037/0022-3514.66.1.37

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science*, *28*, 209–226. http://doi.org/10.1016/j.cogsci.2003.11.004

Martin, D., & Macrae, C. N. (2007). A face with a cue : Exploring the inevitability of person categorization. *European Journal of Social Psychology*, *37*, 806–816. http://doi.org/10.1002/ejsp.445

Mauss, I. B., McCarter, L., Levenson, R. W., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, *5*(2), 175–190. http://doi.org/10.1037/1528-3542.5.2.175

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration

of judgments. *Journal of Personality and Social Psychology*, *35*(4), 250–256. http://doi.org/10.1037/0022-3514.35.4.250

Oosterhof, N. N., & Todorov, A. (2009). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092. http://doi.org/10.1073/pnas.0805664105

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. In *6th IEEE International Conference on Advanced Video and Signal based Surveillance* (pp. 296–301). Italy: IEEE Computer Society. http://doi.org/10.1109/AVSS.2009.58

Ratner, K. G., Dotsch, R., Wigboldus, D. H. J., van Knippenberg, A., & Amodio, D. M. (2014). Visualizing minimal ingroup and outgroup faces: Implications for impressions, attitudes, and behavior. *Journal of Personality and Social Psychology*, *106*(6), 897–911. http://doi.org/10.1037/a0036498

Schönborn, S., Egger, B., Morel-Forster, A., & Vetter, T. (2017). Markov chain monte carlo for automated face image analysis. *International Journal of Computer Vision*, *123*(2), 160–183. http://doi.org/10.1007/s11263-016-0967-5

Silver, M., & Sabini, J. (1978). The perception of envy. *The Perception of Envy*, *41*, 105–117.

Smith, R. H., & Kim, S. H. (2007). Comprehending envy. *Psychological Bulletin*, *133*(1), 46–64. http://doi.org/10.1037/0033-2909.133.1.46

Stolier, R. M., Hehman, E., & Freeman, J. B. (2018). A dynamic structure of social trait space. *Trends in Cognitive Sciences*, 1–4. http://doi.org/10.1016/j.tics.2017.12.003

Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences*, *115*(37), 9210–9215. http://doi.org/10.1073/pnas.1807222115

Susskind, J. M., Lee, D. H., Cusi, A., Feiman, R., Grabski, W., & Anderson, A. K. (2008). Expressing fear enhances sensory acquisition. *Nature Neuroscience*, *11*(7), 843–850.

http://doi.org/10.1038/nn.2138

Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion*, *13*(4), 724–738. http://doi.org/10.1037/a0032335

Todorov, A., Dotsch, R., Wigboldus, D. H. J., & Said, C. P. (2011). Data-driven methods for modeling social perception. *Social and Personality Psychology Compass*, *5*(10), 775–791. http://doi.org/10.1111/j.1751-9004.2011.00389.x

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *43*(2), 161–204. http://doi.org/10.1080/14640749108400966

Van de Ven, N., Zeelenberg, M., & Pieters, R. (2011). Why envy outperforms admiration. *Personality and Social Psychology Bulletin*, *37*(6), 784–795. http://doi.org/10.1177/0146167211400421

Walker, M., & Keller, M. (2019). Beyond attractiveness : A multi-method approach to study enhancement in self- recognition on the Big Two personality dimensions. *Journal of Personality and Social Psychology*, (February). http://doi.org/10.1037/pspa0000157

Walker, M., Schönborn, S., Greifeneder, R., & Vetter, T. (2018). The basel face database: A validated set of photographs reflecting systematic differences in big two and big five personality dimensions. *PLoS ONE*, *13*(3), 1–20. http://doi.org/10.1371/journal.pone.0193190

Walker, M., & Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision*, *9*(11), 1–13. http://doi.org/10.1167/9.11.12

Walker, M., & Vetter, T. (2016). Changing the personality of a face: Perceived Big Two and Big Five personality factors modeled in real photographs. *Journal of Personality and Social Psychology*, *110*(4), 609–624. http://doi.org/10.1037/pspp0000064

Westfall, J. (2016). *PANGEA: Power ANalysis for GEneral Anova designs*. Retrieved from http://jakewestfall.org/pangea/

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychologoical Science*, *17*(7), 592–598.

Young, A. I., Ratner, K. G., & Fazio, R. H. (2014). Political attitudes bias the mental representation of a presidential candidate's face. *Psychological Science*, *25*(2), 503–510. http://doi.org/10.1177/0956797613510717

Appendix C

Walker, M. & **Keller, M. D.** (2019). Beyond attractiveness: A multimethod approach to study enhancement in self-recognition on the Big Two personality dimensions. *Journal of Personality and Social Psychology*. Advance online publication. doi:10.1037/pspa0000157

Beyond attractiveness: A multi-method approach to study enhancement in self-recognition on the Big Two personality dimensions

Mirella Walker and Matthias Keller

University of Basel

Author Note

Correspondence concerning this article should be addressed to Mirella Walker, Department of Social Psychology, University of Basel, Missionsstrasse 64a, CH-4055 Basel, Switzerland, E-Mail: mirella.walker@gmail.com

Draft version: February 6th, 2019

To be cited as: Walker, M. & Keller, M. (2019). Beyond attractiveness: A multi-method approach to study enhancement in self-recognition on the Big Two personality dimensions. *Journal of Personality and Social Psychology.* Manuscript in press. doi: 10.1037/pspa0000157

Abstract

Self-enhancement refers to the phenomenon that individuals tend to have unrealistically positive self-views. Traditional measures of self-enhancement typically imply self-evaluations and reference values, such as evaluations by others or evaluations of the average other. Comparing individuals' self-evaluations with such reference values, however, bears risks. It is not evident that the reference values are more accurate than the self-evaluations and it is not possible to distinguish self-enhancers from individuals who are indeed superior to others. Here, we present two novel methods to measure self-enhancement that circumvent these problems by using participants' own faces as reference values. In Study 1 we systematically manipulate facial characteristics that have previously been found to impact perceptions of attractiveness, likeability, and the Big Two personality dimensions in participants' faces and ask them to recognize themselves. In Study 2 we use a novel approach to apply random noise patterns to participants' faces and ask them to indicate in which version they recognize themselves more. Aggregating these random noise patterns reveals the direction of self-recognition in a more bottom-up, data-driven way. Across both studies we find evidence for self-enhancement regarding attractiveness, likeability, and the Big Two personality dimensions.

Keywords*: self-recognition, self-enhancement, statistical face models, attractiveness, personality*

Beyond attractiveness: A multi-method approach to study enhancement in self-recognition on the Big Two personality dimensions

Healthy individuals' self-perceptions are not always perfectly accurate. They, for example, think that they are more cooperative and intelligent than others (Alicke, 1985), expect a brighter future for themselves than for others (Shepperd, Klein, Waters, & Weinstein, 2013; for a review of the better-than-average effect, see Alicke & Govorun, 2005), or see themselves in more flattering terms than they are seen by others (Lewinsohn, Mischel, Chaplin, & Barton, 1980). Because all these findings have in common that individuals' evaluations of themselves are more favorable than the reference values they are compared to (i.e., their *evaluations of others* or *evaluations by others*), they have been considered as instances of self-enhancement, the phenomenon that individuals tend to have unrealistically positive self-views (Alicke & Sedikides, 2009; Taylor & Brown, 1988).

This reasoning implies that the reference values are the more accurate measure (of a true self-enhancement score) than the self-evaluations. Otherwise, the more favorable self-evaluations compared to the reference values would not necessarily imply that individuals' self-evaluations are inflated. It could also imply that the reference values are deflated. And there are several reasons to believe that the reference values are as susceptible to distortion as self-evaluations are. Because evaluations of others and evaluations by others are both evaluations from a third person perspective, their accuracy depends on the relevance, availability, detection, and utilization of behavioral cues (Funder, 1995). The individual has much more available information about him- or herself than any other person does and, therefore, there is reason to believe that self-evaluations might be more accurate than evaluations by or of others. He or she, for example, knows about his or her actions in a variety of different contexts over time (Krueger,

Ham, & Linford, 1996; Malle & Pearce, 2001) and is also familiar with his or her inner states, such as his or her intentions (Funder, 1995). Whereas the individual might have an advantage regarding the availability of information about him- or herself that has been downplayed in research on self-enhancement (Krueger & Wright, 2011), he or she might not be motivated to see and present him- or herself in the most objective but rather in the most positive light (Paulhus, 1984), which might add a bias to his or her judgment. Following this same logic, however, one could also argue that individuals might be motivated to evaluate others more negatively than they actually are, because this results in a favorable self-other comparison. Thus, the motivation to view oneself in a positive light might limit the accuracy of judgments of and by others (i.e., the reference value) just as much as they limit the accuracy of self-evaluations.

Imagine Lisa, who evaluates herself as a) more competent than she evaluates the average other person and as b) more competent than she is evaluated by others. In this case it might be possible that Lisa is correct in both her evaluations of herself and the average other person, but the others underestimate her competence. However, it might also be possible that Lisa overestimates her own competence, whereas others evaluate her competence quite accurately. These classical discrepancy measures fail to separate judgment error from bias (Heck & Krueger, 2015; for a critical review of these classical discrepancy measures, see also Krueger, Heck, & Asendorpf, 2017). The same pattern of results can describe a self-enhancer who erroneously believes to be superior to others but is evaluated by others accurately or a person who is indeed superior to but at the same time underestimated by others. These measures thus "conflate a false sense of superiority with true superiority" (Heck & Krueger, 2015, p. 1003). To solve this problem, Heck and Krueger (2015) recently developed an alternative measure of self-enhancement (i.e., the projection index) that still relies on self-evaluations and evaluations of

others (i.e., the average other) and allows separation of self-enhancement effects in a rationally justified self-enhancement bias and an error. Measuring self-judgments, other-judgments, and desirability of several traits, they found that self-judgments predict other-judgments to a small extent, that is, individuals likely use their self-judgments to estimate what the average other might be like. In other words, they project their self-image on the average other. Based on this finding they calculated a projection-index, that is, a projection-based prediction of the other-judgments. Regressing desirability scores separately on self-judgments, projection-based predictions of other-judgments, and other-judgments allowed them to separate a rational component of self-enhancement effects, or a self-enhancement bias (i.e., the degree to which desirability better predicts self-judgments than projection-based other-judgments) from a non-rational component, or self-enhancement error (i.e., the degree to which desirability better predicts projection-based other-judgments than observed other-judgments).

In this article we present two alternative methods to measure self-enhancement that do not necessitate asking participants to either evaluate themselves or others and do not make the rating dimension explicit. These methods measure the accuracy of self-recognition regarding various dimensions and allow the comparison of effects of self-enhancement across different groups of individuals or across individuals in different situational contexts, providing information on inter-group, inter-individual, and intra-individual differences in self-enhancement.

The first of these novel methods builds on the innovative approach of Epley and Whitchurch (2008). They measured participants' self-associations by letting them identify their own face among a set of manipulated versions of their face, half of which were reduced in attractiveness, and half were enhanced (henceforth referred to as multiples self-recognition task).

The attractiveness-enhanced versions were made by morphing each participant's face with another face that was created by averaging dozens of faces of the same gender. This resulted in a more average-looking version of the participant's face, which is generally perceived as more attractive (Langlois & Roggman, 1990). The attractiveness-reduced versions were made by morphing each participant's face with the face of a person of the same gender who suffered from a craniofacial syndrome. This procedure resulted in asymmetric versions of the participants' faces, which are perceived to look less attractive. Results showed that participants in this self-recognition task opted for the more attractive versions of themselves, supporting the self-enhancement hypothesis with regard to attractiveness. Moreover, they showed that participants who are told to identify their own face among a set of other persons' faces as quickly as possible are faster in doing so for portraits with enhanced attractiveness than for portraits with reduced attractiveness (henceforth referred to as reaction time self-recognition task). Results from these two different tasks revealed that individuals see themselves as more attractive than they actually are and thus provide evidence for self-enhancement regarding attractiveness (Epley & Whitchurch, 2008).

This paradigm that measures the accuracy of self-recognition and the direction of distortion is a great achievement since it allows for objective comparison of participants' self-recognition of their faces with their actual faces without involving any external standard of comparison. Moreover, given that face recognition is a highly automatic process (Liu, Harris, & Kanwisher, 2002) and participants are unaware of the objective of the task, it is not likely that this measure is susceptible to correction strategies, for example, due to social desirability. One might argue that a disadvantage of this method is that it is restricted to measuring self-enhancement with regard to facial traits, while people do not only have inaccurate perceptions

regarding their own facial traits, such as their attractiveness, but also regarding diverse personality traits (John & Robins, 1994). Reviewing literature on personality and biases in self-perception assessed via self-reports, Paulhus and John (1998) identified two major domains in which individuals have been shown to hold unrealistically positive views of themselves, namely the domains of agency and communion. Whereas unrealistically positive self-perceptions regarding agency trace back to an egoistic bias or a tendency to overstate one's own status, unrealistically positive self-perceptions regarding communion trace back to a moralistic bias or a tendency to understate socially deviant impulses. The authors argue that unconscious self-deceptive processes (as opposed to more conscious forms of impression management) might be at work in both domains, which renders them both potential candidates for our novel methods to measure self-enhancement.

In what follows, we discuss why we believe that the approach to measure self-enhancement applied by Epley and Whitchurch (2008) can be transferred from the domain of attractiveness to the domain of personality (i.e., agency and communion). To do so, we first decompose the successful measurement of attractiveness enhancement into four preconditions and subsequently show that all these preconditions can be met for different personality dimensions as well. The four critical preconditions of successful measurement of attractiveness enhancement in the work of Epley and Whitchurch (2008) were that a) individuals ascribe attractiveness based on faces and b) different individuals do so similarly (i.e., there is a socially shared facial attractiveness stereotype). Moreover, the authors c) had knowledge about the facial information corresponding to this attractiveness stereotype and d) were able to manipulate this information in the faces of the participants in a realistic-looking way.

Individuals, however, do not only ascribe attractiveness based on faces, but they also a) spontaneously and rapidly ascribe personality traits based on faces (Liu et al., 2002; Willis & Todorov, 2006) and b) they highly agree when doing so (Albright, Kenny, & Malloy, 1988; Oosterhof & Todorov, 2008; Walker & Vetter, 2016), even if they have different cultural backgrounds (i.e., Western vs. Asian; Walker, Jiang, Vetter, & Sczesny, 2011). Within the last ten years, different methods have been successfully developed c) to identify and systematically describe these socially shared facial stereotypes of different personality dimensions, such as trustworthiness and dominance (Oosterhof & Todorov, 2008), or the Big Two and the Big Five in faces (Walker & Vetter, 2009, 2016). Combining the systematic description of these facial personality stereotypes with up-to-date computer graphics techniques d) allows the manipulation of this information in the faces of participants in a realistic-looking way (Walker & Vetter, 2016). Because all four of these critical preconditions are met for the domain of personality, we believe that combining this image manipulation technique with the paradigm of Epley and Whitchurch (2008) allows measurement of self-enhancement regarding different personality dimensions. We assume that the portraits participants recognize as their own portraits reflect how they see themselves – not only regarding attractiveness, but also regarding personality.

The facial stereotypes of the Big Two personality dimensions (i.e., agency and communion), as well as likeability and attractiveness, have been previously identified and defined using the idea of a face space (Valentine, 1991). The statistical face space is derived from the analysis of scans of real faces (Paysan, Knothe, Amberg, Romdhani, & Vetter, 2009). The dimensions of the space are defined by the information on which these faces maximally vary. Every individual face is represented as a point in this space. Similar faces lie closer to each other in that space, whereas more diverse faces lie farther apart from each other. Vectors can be

used to describe the difference between any two faces in this space by specifying their deviation from each other on the various dimensions of the space. Collecting Big Two, likeability, and attractiveness judgments for these faces allows computation of vectors with maximum variability regarding the Big Two, likeability, and attractiveness, respectively. These vectors can be applied to novel photographs of faces resulting in realistic-looking versions of them with different levels of perceived personality, likeability, and attractiveness (Walker & Vetter, 2016).

This face modeling approach does not only have the advantage that it can be applied to dimensions that are not primarily facial dimensions, but it also allows a systematic manipulation of attractiveness, independent of any other dimension that describes variations between different individuals. Epley and Whitchurch (2008) reduced facial attractiveness by morphing participants' faces with the facial characteristics of other individuals suffering from a craniofacial syndrome and they enhanced facial attractiveness by morphing participants' faces with a composite face. So, even with no enhancement motive at work, the task to correctly recognize oneself from a set of faces with different degrees of facial distortions (i.e., the lower range of the attractiveness scale) might be easier than the task to recognize oneself from a set of faces with different degrees of one's own facial features, which is the case when one's own face is morphed with a more average-looking face (i.e., the upper range of the attractiveness scale). So, one might argue that, in this case, attractiveness is somehow confounded with task difficulty. This potential confound can be circumvented by using the aforementioned approach to manipulate facial attractiveness. This approach does not involve morphing between different facial identities and it allows incorporation of all aspects of attractiveness (e.g., sexual dimorphism, symmetry, and averageness) as perceived by others instead of restricting it to symmetry.

One might argue that due to it's theory-driven nature, this first method to measure self-enhancement only allows finding enhancement effects on dimensions that were a priori defined to be potential dimensions of self-enhancement (e.g., attractiveness). This objection could be met by the second method that measures self-enhancement in a fully data-driven way. This method combines our face modeling approach with the classical reverse correlation technique that overlays 2D base images of (averaged) faces with different random noise patterns (Dotsch, Wigboldus, Langner, & Van Knippenberg, 2008; Kontsevich & Tyler, 2004; Mangini & Biederman, 2004). In the classic reverse correlation approach, participants are presented with many different pairs of faces that are all created using the same base image and they are repeatedly asked to indicate which exemplar of each pair better matches a given description or group (e.g., Moroccans). The pairs only differ regarding the specific random noise pattern (and its negative) that is imposed on the base image. Translating this technique from the realm of 2D to 3D face representations as used in the aforementioned face modeling approaches allows the use of participants' own faces as base faces and application of the same set of random vectors to all participants' faces. Every vector can be added and subtracted from the participants' faces to create pairs of faces lying at equal distances but in opposite directions from the base face. Participants can then be asked to answer the same question (*In which image do you recognize yourself more?*) repeatedly with different pairs of stimuli, which then allows a posteriori extraction of the exact dimension of recognition distortion.

In this paper we aim to investigate whether individuals self-enhance regarding different facial and personality dimensions. To do so, we pursue a multi-method approach to measure self-enhancement by applying three different paradigms in two separate studies. In Study 1 we systematically model portraits of participants' own faces regarding the salience of perceived

attractiveness, likeability, agency, and communion (i.e., theory-driven method). In Study 2 we go beyond this approach by not defining the dimensions of self-enhancement a priori, but by developing individual self-recognition vectors based on a novel reverse correlation approach (i.e., data-driven method). Comparing these self-recognition vectors with different personality vectors allows the direction and degree of self-enhancement to be quantified.

First, we hypothesize that participants are fastest in recognizing their own face among eight portraits of distractor individuals if the valence of their face is most positive (attractiveness-, likeability-, agency-, and communion-enhanced), followed by the original face and the negative version of their face (attractiveness-, likeability-, agency-, and communion-reduced; *reaction time task*, Study 1). Second, we hypothesize that participants recognize themselves in too positive versions of their own faces when they are asked to identify their actual faces among nine versions of their own face that vary regarding their salience of attractiveness, likeability, agency, and communion (*multiples task*, Study 1). Third, we hypothesize that individual self-recognition vectors reflect self-enhancement on both Big Two personality dimensions (*image classification task*, Study 2).

Finally, we aim to investigate the role of self-esteem on self-enhancement. Previous work has shown that there is a positive relation between implicit, but not with explicit self-esteem and self-enhancement, revealing that individuals with low implicit self-esteem less strongly enhance than individuals with high implicit self-esteem (Epley & Whitchurch, 2008). We aim to investigate whether these results replicate with our novel approach to measuring self-enhancement.

Study 1

In Study 1 we use a validated technique to model the faces of our participants so that they look more and less attractive, likeable, agentic, and communal (Walker & Vetter, 2016) and measure whether individuals recognize themselves (faster) in too positive versions of themselves.

Method

Since photographs of participants' faces had to be taken and manipulated to investigate the impact of these manipulations on self-recognition, this study was divided into two sessions. The first session mainly consisted of taking photographs of participants' faces and collecting data about a possible moderator (i.e., explicit self-esteem), whereas the second session consisted of the main experiment.

Participants. A total of 64 student participants (53 female, 11 male) completed this study. Their mean age was 24.03 years (*SD* = 5.18). They were paid an amount equivalent to 15 USD in the local currency for participation. A provisional sample size was estimated based on the studies of Epley and Whitchurch (2008). Because our manipulations are subtler than theirs, we expected effect sizes to be somewhat lower and therefore aimed for a sample size of 40. We then ran a power analysis based on these provisional results[1] and determined the definite sample size accordingly (i.e., n = 62). We aimed for two more participants in case some participants did not finish the second session of the study.

Material. Photographs of all participants' faces were taken with a Canon MV700 camera. The faces from these 2D photographs were then reconstructed in 3D using an analysis-by-synthesis approach in which we linearly combined the 200 faces that the Basel Face Model is

[1] We used the provisional effect size for the second smallest self-enhancement effect (i.e., agency enhancement) in the multiples task (*d* = .32) to estimate the final sample size. The

built upon (Paysan et al., 2009). The resulting 3D estimates of participants' faces were manipulated by applying a successfully validated approach to subtly and systematically enhance and reduce the salience of the four different dimensions in faces. As specified in the introduction, these dimensions were previously determined based on participants' judgments of the faces in the Basel Face Model. Then, we rendered the resulting versions of participants' 3D face estimations back into the original photographs with realistic-looking results (for details, see Walker & Vetter, 2016). Finally, the resulting images were mirrored, based on the assumption that people are more familiar with their mirrored faces than with their faces as shown in photographs (Mita, Dermer, & Knight, 1977; Rhodes, 1986). Here we manipulated perceived attractiveness, likeability, agency, and communion. Likeability is used as an approximation for valence (see, e.g., Singh, 2014). The Big Two dimensions of agency and communion (Wiggins, 1991) are ideal candidates to test our approach to measure self-enhancement for several reasons. They correspond to two important biases in self-perception, namely, an egoistic and a moralistic bias (Paulhus & John, 1998). Moreover, the two dimensions are similar to the two fundamental dimensions of face perception (i.e., dominance and trustworthiness, Oosterhof & Todorov, 2008), they are both perceived to be positive in valence (Suitner & Maass, 2008) and they are theoretically (Abele & Wojciszke, 2007) and statistically independent from each other ($r(151) = -.02$; for more details regarding the development and validation of these vectors, see Walker & Vetter, 2016).

Since attractiveness is a dimension that is highly gender-specific, meaning that different characteristics are perceived to be attractive in male and female faces (i.e., sexual dimorphism, Perret et al., 1998), we applied gender-specific attractiveness-vectors to our participants' faces. Manipulations were very subtle and the resulting portraits were realistic-looking, such that

participants should not have realized that they were presented with manipulated versions of their original portraits. For the reaction time self-recognition task, 3 by 3 matrices of portraits were created, among which one was a manipulated version of the participant's own face (i.e., the target face) and eight were original photographs of other participants that served as distractors. For the multiples self-recognition task we created 3 by 3 matrices of versions of the participant's own face manipulated on the same dimension (e.g., attractiveness). The two faces from the reaction time self-recognition task were the most extreme among them. The remaining seven versions were manipulated to lie at equal distances on that continuum between the two extremes.

To investigate a potential moderator of biased self-recognition, we measured self-esteem with explicit and implicit measures. A German version of Rosenberg's Self-Esteem Scale was used to measure explicit self-esteem (Janich & Boll, 1982). To measure implicit self-esteem, the Name-Letter-Task (NLT; Kitayama & Karasawa, 1997) and a Single-Target IAT (ST-IAT; see e.g., Bluemke & Friese, 2008) were used. To rule out any moderation effects of mood, we measured mood with a German mood scale (Aktuelle Stimmungsskala ASTS; Dalbert, 1992).

Procedure.

Session 1. Participants were first asked to put on a black t-shirt and pull their hair back, so that no parts of the face were covered and clothing was identical. They were then asked to adopt a neutral facial expression and sit on a chair in front of a white background, where they were photographed from the front. Then participants were seated in front of computers and asked to read the instructions and fill in a questionnaire[2]. The questionnaire was a 10-item (e.g., "I think that I possess many strengths") German version of Rosenberg's Self-Esteem Scale (Janich & Boll, 1982). Then, they took part in two short unrelated studies.

[2] Participants of whom we already had taken a portrait for a previous study were sent this questionnaire by email and answered it online.

Session 2. After approximately one week, participants were sent a link to complete part 2 of the study online. Participants were asked to turn off all other electronic equipment besides their computer and not to engage into any other activity during the completion of the online study. Their first task was to complete a self-esteem ST-IAT. In the first practice block, positive and negative terms had to be classified. In the following initial test block, self-relevant words had to be classified along with the positive words. In the second test block, the self-relevant words had to be classified along with the negative words. Presentation and position of the stimuli within a block was random despite the first stimulus per block, which was always identical and treated as a test trial. Each block consisted of 70 trials (i.e., 20 positive, 20 self-relevant, and 30 negative items if the former two had to be classified together and 30 positive, 20 self-relevant, and 20 negative items if the latter two had to be classified together). The proportion of left vs. right key concepts thus was always 4:3 or vice versa (see, e.g., Bluemke & Friese, 2008). Then, participants were asked to indicate their mood on a 16-item mood scale with a 7-point Likert scale (i.e., ASTS; Dalbert, 1992). Next, we measured self-enhancement with regard to attractiveness, likeability, agency, and communion using two different paradigms. The first task was a reaction time self-recognition task. Participants were presented with nine different portraits on one screen positioned in a 3 by 3 matrix. One portrait showed the participant's original face or a slightly manipulated version of it, whereas the other eight portraits showed the original portraits of other participants' faces and served as distractors. Participants had to indicate, as quickly as possible, whether their own portrait was in the left, middle, or right column using the arrow keys. Since every version of the participant's face (i.e., plus/minus attractiveness, likability, agency, communion, and the original) was presented nine times the whole task

consisted of 81 trials presented in a random order. The participant's face appeared at random positions in these 3 by 3 matrices.

In the multiples self-recognition task we measured self-enhancement with regard to attractiveness, likeability, agency, and communion by presenting participants with nine different versions of their portrait on one screen arranged in a 3 by 3 matrix (see Figure 1 for an example). One of these portraits was the original photograph, whereas the others showed the participant with different (both reduced and enhanced) degrees of ascribed attractiveness, for example. Arrangement of stimuli was random. Participants could take as much time as needed to select the face that they perceived as their real face. This procedure was repeated four times in total with screens showing faces with differing degrees of ascribed attractiveness, likeability, agency, and communion. Then, the ASTS (Dalbert, 1992) was presented again to assess mood after measuring the dependent variables. Then, the NLT (Kitayama & Karasawa, 1997) was presented, in which participants had to indicate their attitude towards all letters of the alphabet, in order to assess participants' preference for the letters in their initials as compared to all other letters in the alphabet. Finally, participants were asked to give some demographical information. They were thanked for participation and given a code to receive payment for participation.

Results

Reaction time self-recognition task. We had to omit one participant from these reaction time analyses, because the participant indicated at the end of the study that s/he always had to scroll down to see all nine faces, which is problematic for the dependent variable reaction time. For the remaining participants, we first removed reaction times from those trials in which participants mistakenly chose a column that did not include their own portrait or a variation of it. Then we winsorized reaction time outliers that deviated by more than 3 standard deviations from

the respective average. After these data-preprocessing steps, we averaged reaction times across the nine trials with the same version of participants' faces resulting in nine reaction time scores (i.e., for the original faces and for the versions with enhanced and reduced salience of attractiveness, likeability, agency, and communion). The Kolmogorov-Smirnov test revealed that two out of nine reaction time scores were not normally distributed, $D_{Attr\_pos}(63) = 0.116$, $p = 0.034$ and $D_{Comp\_pos}(63) = 0.141$, $p = 0.003$. Therefore, logarithmic transformations were applied to all nine reaction time scores. To test whether participants showed a self-recognition advantage for positive vs. original vs. negative versions of their faces we ran both linear and quadratic trend analyses (salience of characteristic: enhanced vs. original vs. reduced) with the dependent variable average reaction time. Supporting Hypothesis 1, participants were fastest in recognizing the positive versions, followed by the original versions, followed by the negative versions of their faces for attractiveness, $F_{\text{linear}}(1, 62) = 10.92$, $p = .002$, $\eta_p^2 = .150$, likeability, $F_{\text{linear}}(1, 62) = 7.98$, $p = .006$, $\eta_p^2 = .114$, and agency, $F_{\text{linear}}(1, 62) = 14.51$, $p < .001$, $\eta_p^2 = .190$. For attractiveness and agency, quadratic trends also reached statistical significance; however, these effects were descriptively less pronounced than the linear effects, $F_{\text{attractiveness}}(1, 62) = 9.93$, $p = .003$, $\eta_p^2 = .138$ and $F_{\text{agency}}(1, 62) = 7.24$, $p = .009$, $\eta_p^2 = .105$. In all three cases, effects are driven more strongly by a reaction time disadvantage for the negative version than by a reaction time advantage for the positive version. For communion, neither the linear nor the quadratic trend reached statistical significance (see Table 1 for all means, standard deviations, F, p, and η_p^2-values).

Multiples self-recognition task. All participants were included in this analysis. First, we built a self-recognition score for every participant. Therefore, the nine faces presented in the explicit enhancement task were coded so that the most negative face equaled "1", the most

positive face equaled "9" and the original face equaled "5". We then ran four one-sample t-tests to compare the self-recognition score against a mean of 5 (original face) to test whether participants chose a more positive face when they were asked to identify their original face. Results show that people do choose a significantly too attractive ($M = 5.67$, $SD = 2.44$, $t(63) = 2.20$, $p = .031$, $d = .27$), too likeable ($M = 6.02$, $SD = 2.06$, $t(63) = 3.95$, $p < .001$, $d = .49$), and a too agentic version of their own face ($M = 5.70$, $SD = 2.38$, $t(63) = 2.36$, $p = .021$, $d = .30$). Descriptively, mean values for communion were also above the mean value of 5; however, they missed conventional levels of significance ($M = 5.53$, $SD = 2.43$), $t(63) = 1.75$, $p = .085$, $d = .22$ (see Table 2 for all means, standard deviations, t, p, and d-values).

Interestingly, a correlation analysis between the four dimensions reveals that enhancing with regard to one dimension does not necessarily mean enhancing regarding other dimensions as well. With one exception, self-recognition scores on the four dimensions were independent from each other ($.04 < r\ (62) < .17$, $p_{min} > .176$). Only likeability and communion were highly correlated: Participants who recognized themselves in versions with enhanced communion also recognized themselves in versions with enhanced likeability and vice versa, $r\ (62) = .46$, $p < .001$. If these two conceptually (Abele & Wojciszke, 2014) and statistically similar dimensions are combined into one communion/likeability-scale, then participants select a positively distorted version of their own face on that scale ($M = 5.77$, $SD = 1.92$), $t(63) = 3.22$, $p = .002$, $d = 40$.

Self-esteem, mood, and self-enhancement. To test for moderating effects of self-esteem and mood on self-enhancement, we inspected correlations of the different self-esteem and mood measures a) with the reaction time differences between the negative and the positive versions of participants' portraits in the reaction time task and b) with the self-recognition scores in the multiples task. Because hypothesizing that there is an effect of explicit self-esteem [implicit self-

esteem or mood] on self-enhancement involves running four [eight each] significance tests for the reaction time task and four [eight each] significance tests for the multiples task, we correct alpha levels to 0.013 [0.006 each].

Self-esteem, mood, and self-enhancement in the multiples task. Explicit self-esteem and enhancement scores on the four different dimensions were not significantly correlated, $|r|_{max}$ *(62)* = -.15, p_{min} = .237. Because the two implicit measures of self-esteem, namely NLT[3] and Self-Esteem IAT[4], were hardly correlated, *r (60)* = .23, p = .071, they were considered separately in the analyses. We did not find significant effects of implicit self-esteem or mood on self-recognition, $|r|_{max}$ (60) =.227, p_{min} = .076 and $|r|_{max}$ *(62)* = .284, p = .023, respectively.

Self-esteem, mood, and self-enhancement in the reaction time task. Reaction time difference scores were neither correlated with explicit self-esteem, $|r|_{max}$ (61) = .086, p_{min} = .501, nor with implicit self-esteem, $|r|_{max}$ (61) = .229, p_{min} = .071. None of the correlations between the two mood measures with the reaction time difference scores reached statistical significance, $|r|_{max}$ (61) = .198, p_{min} = .120.

Discussion

The aim of this study was to investigate whether individuals' tendency to self-enhance regarding their attractiveness generalizes to a general valence dimension and to the Big Two personality dimensions when no external reference value is involved and when they do not deliberately judge themselves. Mostly supporting our first hypothesis, results from the reaction time self-recognition task (i.e., self among others) revealed that participants do not only show a reaction time advantage when identifying their face if a positive and purely face-based

[3] NLT scores were calculated according to Albers, Rotteveel and Dijksterhuis (2009). Please note that NLT data of two participants are missing.
[4] IAT data were analyzed according to the improved algorithm developed by Greenwald, Nosek, and Banaji (2003).

dimension (i.e., attractiveness) is enhanced, but also if a general valence dimension (i.e., likeability) and one of the two fundamental personality dimensions, namely agency, is enhanced. However, they do not show this self-identification advantage if the other fundamental personality dimension, namely communion, is enhanced.

Supporting our second hypothesis, results of the multiples self-recognition task (i.e., different versions of the self) revealed that participants show systematic recognition distortions: They select versions of their faces that look more attractive, more likeable, and more agentic than their actual faces. The pattern looks similar for communion. Correlations between enhancement scores on the different dimensions revealed a large correlation between communion- and likeability-enhancement, signaling that these two dimensions are not only conceptually (Abele & Wojciszke, 2014), but also statistically similar. It seems likely that the results show an over-estimation of one effect (i.e., likeability) and an under-estimation of the other (i.e., communion) with the true effect lying in between. Indeed, a combined likeability/communion-score that contains less noise reveals a significant enhancement on this likeability/communion dimension.

One explanation for these somewhat mixed findings regarding the communion dimension might be that there are systematic differences in enhancement behavior between communion and the other three dimensions. Participants might show a self-recognition advantage only for the dimensions that are profitable from the first-person perspective. It is easier for us to achieve our goals if we are active, decisive, confident, and can withstand pressure easily (i.e., agentic). Therefore, agency is highly profitable for the individual, but not necessarily for others (Abele & Wojciszke, 2007). In contrast, it is easier to get along with others who are friendly, empathetic,

understanding, and cordial (i.e., communal). Therefore, communion is highly profitable for others, but not necessarily for the individual (Abele & Wojciszke, 2007).

Another explanation might be that the unclear findings for communion are a methodological artifact. There are both theoretical and methodological reasons to assume so: First of all, being communal is not always a burden. There are contexts in which communal traits are not only beneficial for others, but also for the person him- or herself (e.g., in building social networks). Second, the pattern of results for communion has been shown to resemble the pattern of results for likeability. The significant effect for the combined likeability/communion-score in the multiples self-recognition task might reflect a more reliable finding then the results for the two separate scores. Third, validation data for the agency and the communion vectors in faces of other individuals presented previously (Walker & Vetter, 2016) revealed that individuals descriptively more easily detect agency than communion manipulations in faces.

Interestingly, the stronger effects for agency than for communion found here contravene previous work providing evidence for stronger enhancement effects for morality than for intelligence (i.e., the Muhammad Ali Effect, Allison, Messick, & Goethals, 1989), two dimensions highly overlapping with communion and agency. One reason for this difference in previous work is that there are more objective criteria to measure intellectual than moral abilities and individuals enhance more if such objective criteria are missing (Alicke & Govorun, 2005). If individuals, however, do not need to make explicit evaluations – neither of themselves nor of others – and the dimension under investigation is not obvious, as in our present study, this self-enhancement advantage of the moral domain is likely to disappear.

Epley and Whitchurch (2008) show that implicit self-esteem moderates self-enhancement effects. We did not find any evidence for a moderation of self-enhancement via implicit or

explicit self-esteem or mood. At this point, we can only speculate why our data does not replicate the moderation effect found in this previous work. However, because we did not directly replicate Epley and Whitchurch's (2008) study but, for example, used German versions of the self-esteem measures and an updated method to measure self-enhancement via four dimensions with more realistic-looking stimuli, this divergence in our findings should probably not be given too much weight. Because the two rather extensive implicit self-enhancement measures were hardly correlated with each other and neither of them showed correlations with any of the eight self-recognition measures, we will refrain from measuring implicit self-esteem in Study 2.

Study 2

To address whether individuals selectively enhance regarding some personality dimensions (e.g., agency) but not others (e.g., communion) or whether the difference between agency and communion was a methodological artifact of Study 1, we focus on agency and communion in Study 2, using a different approach to measure self-enhancement. Again, this approach does not involve any reference value but the participants' own faces. In contrast to the approach we used in Study 1, we do not a priori define the dimensions on which we expect effects of self-enhancement. Instead of applying previously developed vectors that systematically enhance the salience of specific personality dimensions in faces, we apply random noise to these faces and let participants solve a forced choice task in which one face shows the participant with a random noise pattern, whereas the other face represents themselves with the negative version of that random noise pattern. This approach has the advantage that it allows for the generation of classification images that reflect participants' mental representations of themselves without making any prior assumptions about them. In other words, it allows us to measure self-recognition in a fully data-driven way. The resulting dimensions of enhancement can finally be

compared to previously developed personality dimensions. The reverse correlation approach was initially developed by Kontsevich and Tyler (2004) and Mangini and Biederman (2004) and has been extensively applied to measure internal mental representations about various groups (e.g., Moroccan; Dotsch et al., 2008; nurse and manager; Imhoff, Woelki, Hanke, & Dotsch, 2013; welfare recipients; Brown-Iannuzzi, Dotsch, Cooley, & Payne, 2017). We refined this approach by integrating it into the face space approach and adding up-to-date image manipulation techniques (Walker & Vetter, 2016). A first application of this combined approach was successful in demonstrating that individuals' beliefs about the association of different personality dimensions translates into how these personality dimensions are inferred from faces (Stolier, Hehman, Keller, Walker, & Freeman, 2018). Here we bring this combined approach to a new level by applying it to portraits of our participants allowing us to measure distortions in self-recognition.

Method

This study was part of a larger set of studies, which all involved manipulated versions of participants' portraits. Therefore, again, this study was comprised of two sessions. Relevant for the present study, participants in the first session were photographed and they filled out a questionnaire measuring a possible moderator (i.e., explicit self-esteem) and the demographic variables, whereas the second session consisted of the main experiment.

Participants. A total of 113 student participants (92 female, 21 male) completed the study for course credit. Their mean age was 21.59 years (SD = 5.22). Because this is the first study to use a combined reverse correlation / face space approach measuring self-enhancement, we could not rely on previously obtained effects for power calculations. Therefore, we aimed to recruit between 100 and 120 participants. This number is considerably higher than the number

usually recruited in reverse correlation studies with a similar dimension of trials (e.g., Dotsch & Todorov, 2012), because we wanted to make sure we would be able to detect even small effects if these effects exist.

Material. Photographs of all participants' faces were taken with a Canon EOS 70D. After a few preprocessing steps (for details see Walker, Schönborn, Greifeneder, & Vetter, 2018), the faces on these photographs were reconstructed in 3D. Instead of applying previously developed vectors that correspond to the perception of specific (personality) dimensions, we created 400 random vectors in the multidimensional face space, with 200 random vectors varying regarding the first 50 and thus most meaningful shape components, and 200 randomly varying regarding the first 50 color components. To do so, we first generated 4 blocks of 100 random vectors each. Because every block consisted of 100 random vectors with 50 principal components and because we wanted to have a similar variance regarding all 50 principal components, we created 50 random distributions of 100 numbers per block with the restriction that these 100 numbers always had a mean of 0 and a standard deviation of 0.3. The first random distribution was then used to define the eigenvalues of the first principal component for all 100 vectors in that block, the second random distribution was used to define the eigenvalues of the second principal component for all 100 vectors in that block, and so on for all 50 distributions or principal components and for all 4 blocks.

These random vectors were then individually applied to the 3D estimations of the participants' faces. This procedure resulted in 400 pairs of faces for every participant: In one exemplar per pair, the random vector was applied to enhance the respective random information (i.e., vector addition), whereas in the other, the random vector was applied to reduce the respective random information (i.e., vector subtraction). Finally, all stimuli were mirrored to

account for the fact that individuals are more familiar with their mirrored face than with their face from the outside perspective (Mita et al., 1977; Rhodes, 1986). Figure 2 shows some exemplar pairs for one female and one male participant.

To investigate self-esteem as a potential moderator for self-enhancement, we again used a German version of Rosenberg's Self-Esteem Scale (Janich & Boll, 1982).

Procedure.

Session 1. As in Study 1, participants were first asked to put on a black t-shirt and pull their hair back, so that no parts of the face were covered and clothing was identical. They were then asked to adopt a neutral facial expression and sit on a chair in front of a white background, where they were photographed from the front. Next, participants were seated in front of computers and asked to read the instructions and fill in a questionnaire. The questionnaire was a 10-item (e.g., "I think that I possess many strengths") German version of Rosenberg's Self-Esteem Scale (Janich & Boll, 1982). Finally, participants were asked to give some demographical information.

Session 2. After approximately one week, participants came back to the lab to complete the study on a computer. Participants were told that they would be repeatedly presented with two variations of their portrait on every screen and they had to choose the version in which they recognized themselves more, that some trials might be more difficult than others, and that their task was just to spontaneously indicate their answer without deliberating on their decision. Then they were presented with the first block of 100 shape trials. This block started with a fixation cross (0.5 seconds) that was followed by the simultaneous presentation of the heading *"In which image do you recognize yourself more?"* and the two portraits manipulated with the first random vector. They could answer by clicking the left or the right arrow key. Then the second fixation

cross appeared and so on. After the first 100 trials, participants were told that the first out of four blocks was finished and that they could take a moment to relax before starting with the second block. The second block consisted of 100 new shape trials. The third and the fourth block consisted of 100 color trials each. The order of the trials within the blocks was completely random. After the fourth block participants were thanked and debriefed.

Results

Preliminary work. To investigate whether individuals self-enhance on both Big Two personality dimensions, agency and communion, we generated individuals' self-recognition vectors and developed male and female agency and communion vectors to compare them to. Figure 3 visualizes the steps of this procedure.

Generation and visualization of self-recognition vectors. First, we created a self-recognition vector individually for every participant by averaging the 200 shape and the 200 color vectors that corresponded to the faces they selected. These individual self-recognition vectors thus contain the facial information participants regard as characteristic for themselves and point in the respective directions in the face space. Applying these vectors to the original portraits of the participants visualizes how they see themselves (see upper left part of Figure 3 for visualizations of fictitious[5] individual self-recognition vectors). Then we averaged these individual enhancement vectors separately for all female and all male participants. These averaged two self-enhancement vectors contain the facial information that females and males regard as characteristic for themselves and point in the respective direction in the face space. Applying these vectors to the average female and male face from the Basel Face Model (Paysan et al., 2009) visualizes how females and males on average see themselves (see bottom part of

[5] Please note that for privacy reasons we did not apply the self-recognition vectors of these specific participants but of other same-gender participants.

Figure 3 for a visualization of the male self-recognition vector). More detailed visualizations of the female and the male self-recognition vectors are presented in the first three rows of Figure 4. The first and the second face in the same row represent the average female face of the Basel Face Model with reduced and enhanced levels of the female enhancement vector, while the third and the fourth face in the same row represent the average male face of the Basel Face Model with reduced and enhanced levels of the male enhancement vector. The first and the second row visualize the color and the shape information separately, while the third row visualizes the same information in combination. Informal visual inspection of these faces reveals that both females and males recognize themselves more in faces that are smaller, especially in the lower part of the face, have smaller mouths and chins, lighter, but bigger eyes and a darker skin tone around the mouths than their own faces. However, there are also differences between the female and the male self-recognition vectors. Females tend to recognize themselves more in faces that have darker, more strongly curved eyebrows and an overall darker skin tone, whereas males tend to recognize themselves in faces that have lighter eyebrows and an overall lighter skin tone.

Development of agency and communion vectors. To generate agency and communion vectors we performed four image classification tasks (i.e., agency female, agency male, communion female, and communion male). We collected data from 239 participants (109 female, 129 male, 1 participant did not indicate his or her gender) with a mean age of 36.51 years (*SD* = 11.74). To generate the stimuli for the male [female] image classification tasks, we applied the same random vectors as in Study 2 to a morph consisting of 100 male [female] faces from the Basel Face Model (Paysan et al., 2009). As in Study 2, we created 400 pairs of faces resulting in 200 shape and 200 color trials. However, each participant only completed 100 shape and 100 color trials. Again, we always presented two faces on the same page and the task for the

participants was to indicate which of the two faces looks more competent, efficient and competitive (i.e., agency condition) or more well-intentioned, trustworthy and sincere (i.e., communion condition). Participants' answers were averaged and used to extract an agency and a communion vector separately for male and female faces. These agency and communion vectors thus contain the facial information participants regard as characteristic for competent, efficient and competitive individuals and well-intentioned, trustworthy and sincere individuals and point in the respective direction in the face space. Applying these vectors to the average male and female face from the Basel Face Model (Paysan et al., 2009) visualizes the facial characteristics perceived as signaling an agentic and a communal personality (see upper right part of Figure 3 for visualizations of the male agency and communion vectors). Validation data for these vectors with 59 independent participants (25 female, 33 male, 1 participant did not indicate his or her gender; M_{age} = 37.64 years, SD = 12.48) and independent faces show that both vectors successfully change the respective personality judgments as reflected in linear trends for agency, $t(31.03) = 2.23$, $p = .03$, and communion, $t(24.12) = 3.96$, $p < .001$ (Keller, Reutner, & Walker, 2017).

Main analyses. Comparing participants' self-recognition vectors with the respective agency and communion vectors (i.e., male [female] vectors for male [female] participants) as visualized in the bottom part of Figure 3 allowed us to test whether participants select their own face if facial information signaling agency and communion is enhanced or, in other words, if participants self-enhance regarding agency and communion. Therefore, we analyzed the data as follows: For all males [females], we correlated the 400 random vectors with the male-specific [female-specific] agency vector and the male-specific [female-specific] communion vector. The absolute values of the correlation coefficients indicate how much each random vector is related

to the dimension in question (i.e., agency or communion). For every participant and every trial, we then recoded the choice variable into two new variables, namely choice of agentic version (0 = no, 1 = yes) and choice of communal version (0 = no, 1 = yes), based on whether the participant chose the face that was positively or negatively correlated with the agency or the communion vector, respectively. If a participant in a specific trial chose the face that was negatively correlated with agency [communion], then the new variable choice of agentic [communal] version was coded 0. If the participant chose the face that was positively correlated with agency [communion], then the new variable choice of agentic [communal] version was coded 1. We then analyzed whether the magnitude of the correlation coefficient predicts these choice variables, which would suggest that individuals do self-enhance regarding these dimensions. If individuals, for example, do self-enhance with regard to agency, then the more strongly a random vector is correlated with the agency vector, the more likely individuals should be to select the more agentic-looking version of these two faces. We performed these analyses using the *glmer* function in *R* (R Core Team, 2017) package *lme4* (Bates, Maechler, Bolker, & Walker, 2014) with random intercepts for participants and face pairs. Supporting our third hypothesis, results revealed that the more strongly a face showed information related to agency, the more likely participants were to select that face (see Table 3 for betas, standard errors, *Z*- and *p*-values). Similarly, the more strongly a face showed information related to communion, the more likely participants were to select that face (see Table 3 for betas, standard errors, *Z*- and *p*-values).

Gender, self-esteem, and self-enhancement. Because hypothesizing that there is an effect of gender or explicit self-esteem on self-enhancement involves running two significance

tests each (i.e., one for agency and one for communion), we correct alpha levels for both potential moderators to 0.025.

In order to investigate whether these effects of self-enhancement are moderated by the gender of participants, we ran two additional glmer-analyses adding a second fixed factor, participant gender, and the interaction between participant gender and the correlation between the random vectors and the agency [communion] vector. Results revealed that males and females differ in their tendency to choose a face that signals high levels of agency; $\beta = 1.56$, $SE = 0.14$, $Z = 11.04$, $p < .001$, whereas they do not differ in their tendency to choose a face that signals high levels of communion; $\beta = .07$, $SE = 0.17$, $Z = 0.43$, $p = .671$. Figure 5 reveals that the higher the correlation between a random vector and agency or communion, the more likely it is that both males and females will select the face of the respective pair that is positively correlated with agency and communion. For both dimensions this tendency seems to be stronger for males than for females. However, the difference between males and females only reaches statistical significance for agency.

To analyze the impact of self-esteem on self-enhancement we first specified the degree to which a participant's individual self-recognition vector was correlated with the gender-specific agency and communion vector, respectively. We then analyzed whether self-esteem can predict the degree to which a person's individual self-recognition vector correlates with the agency and communion vectors or, in other words, a person's tendency to select a face that looks agentic or communal using the lm function in R (R Core Team, 2017). We found that participants with low explicit self-esteem more strongly tend to self-enhance. They are more likely to select the face with higher levels of agency and communion than persons with high explicit self-esteem (see Table 4 for betas, standard errors, *t*- and *p*-values). These findings support the notion that self-

enhancement effects are the result of a motivational process to have a positive self-view. The more this self-view is chronically positive as indicated by the self-esteem measure, the less the individual self-enhances in this task. In other words, the more positive a person perceives him- or herself to be in the classical sense, the less positive he or she perceives him- or herself in the figurative sense.

Discussion

The aim of Study 2 was to figure out whether individuals selectively enhance on personality dimensions that are profitable for themselves or whether they generally enhance on positive personality dimensions. Whereas in Study 1 we a priori defined the dimensions on which we expected participants to enhance, in Study 2 we presented participants with faces that randomly varied on various facial dimensions at the same time (e.g., fullness, mouth shape, brightness, contrast) to allow them to identify the faces that best represented themselves. This approach allows us to create an individual enhancement vector for every participant as well as collapsed over all participants that can be visualized and interpreted. Moreover, these extracted individual or global enhancement vectors can be compared to personality vectors to investigate the kind of enhancement they reflect.

Supporting Hypothesis 3, both visual and statistical inspection of our data reveal that participants enhance regarding both Big Two dimensions. They are more likely to select the agentic or the communal face from a pair when the respective random vector has a higher correlation with agency or communion. This effect is moderated by participant gender. Males tend to self-enhance more strongly than females, especially regarding agency, the dimension that is perceived as stereotypically masculine (Abele, 2003). Moreover, the results of Study 2 show a negative correlation between explicit self-esteem and self-enhancement. This might suggest that

self-enhancement serves a compensatory function. Individuals with low explicit self-esteem have a stronger motivation to see themselves in a positive light (Brown, 2012) and thus to self-enhance regarding the positive Big Two dimensions than individuals with high explicit self-esteem. Individuals who already see themselves in a positive light (high self-esteem) might have a weaker motivation to reach that goal than individuals who do not yet see themselves in a positive light (i.e., have low self-esteem). However, due to the correlational nature of these data and the fact that previous research finds positive rather than negative correlations of self-esteem with self-enhancement (Epley & Whitchurch, 2008; Heck & Krueger, 2015; Taylor & Brown, 1988), the conclusion that self-enhancement serves a compensatory function should be interpreted with caution.

Methodological Considerations

Internal Validity

The two methods we applied to systematically manipulate perceived personality in faces (i.e., theory-driven method, Study 1) and to extract individuals' mental representations of themselves from random noise (i.e., data-driven method, Study 2) have both been validated before (Keller et al., 2017; Walker & Vetter, 2016). We tested our hypotheses using three different paradigms (i.e., reaction time self-recognition, multiples self-recognition, and an image classification task) and measures (i.e., reaction times, recognition distortion, self-resemblance). In the reaction time self-recognition task, participants were repeatedly (i.e., 81 times) asked to identify their own portrait among the portraits of others and we measured their reaction times. In the multiples self-recognition task, they were asked to select their real portrait from a series of portraits with slight variations and we measured the degree of recognition distortion (deviation from the original portrait). In the image classification task, they were repeatedly (i.e., 400 times) asked to choose

one of two versions in which they see themselves more. In contrast to the multiples self-recognition task, participants in this reverse correlation image classification task were never presented with their original portrait (i.e., there were no right or wrong answers).

Sample Size and Diversity

Aiming for a power of .80 in Study 1 and expecting our more subtle manipulations to have a somewhat weaker effect, we considerably enhanced both sample size and number of trials in the reaction time self-recognition task as compared to the studies by Epley and Whitchurch (2008). Based on the data we initially aimed for (i.e., N = 40), we calculated a power analysis and determined the definite sample size (i.e., N = 64) accordingly. Because there are no studies measuring self-recognition with a reverse correlation approach, in Study 2 we oriented ourselves towards other reverse correlation studies (e.g., Dotsch & Todorov, 2012) and enhanced sample size and number of trials in order to make sure to be able to detect even small effects of self-enhancement. Because participants had to come to the lab (even twice in Study 2) to take part in our studies, we had to rely on a pool of undergraduate psychology students in both studies. Therefore, the samples are relatively limited regarding age, they are predominantly female, Western, and well educated. In Study 2, however, there were enough male participants to investigate gender differences. Interestingly, although gender is a critical variable with respect to (self-) ascriptions of agency and communion (i.e., agency is regarded as a stereotypically masculine dimension, whereas communion is regarded as a stereotypically feminine dimension; Suitner & Maass, 2008), we found differences in the degree, but not in the pattern of self-enhancement in males and females. Both females and males significantly enhanced regarding the stereotypically male and female dimension. The fact that two social groups – stereotypically regarded as strongly deviating on the dimensions investigated in these studies – show the same

pattern of results might provide some evidence that the effects of self-enhancement generalize at least to some degree across samples. However, more research focusing on inter-group similarities and differences is needed to confirm this assumption.

General Discussion

In this paper we aimed to investigate whether the finding that individuals have too positive views of themselves with regard to attractiveness replicates and generalizes from the mere facial to the personality domain when only portraits are used to measure self-enhancement. In two studies we investigated the direction and degree of self-enhancement using three different paradigms (reaction time self-recognition, multiples self-recognition, and image classification), different stimulus material (enhancing/reducing specific personality dimensions in participants' faces and enhancing/reducing random dimensions in participants' faces), and different data analytic strategies. Replicating and extending the findings of Epley and Whitchurch (2008) from a facial dimension (i.e., attractiveness) to global valence (i.e., likeability) and personality dimensions (i.e., agency and communion), we found systematic self-enhancement effects. For communion we found significant self-enhancement only in Study 2. Descriptively, the results also show a tendency for communion-enhancement in Study 1. The communion and likeability vectors point in very similar directions in face space, which renders it unlikely that participants enhance regarding one of these dimensions, but not the other. Therefore, we assume that the communion-enhancement was underestimated in Study 1. The finding that there is significant self-enhancement on the combined likeability/communion-score supports this argument. Indeed, by using random vectors in Study 2 we found similar self-enhancement effects for agency and communion. Getting rid of any external standard of comparison leads to enhancement effects on both fundamental personality dimensions.

In Study 2 we also found evidence that explicit self-esteem is systematically involved in self-enhancement. Participants with low self-esteem are more likely to self-enhance on both dimensions (i.e., agency and communion) than participants with high self-esteem. These results suggest that there is a compensatory relation between self-esteem and self-enhancement: The more positive a person perceives him- or herself to be in the classical sense, the less positive he or she perceives him- or herself in the figurative sense. This reverse pattern between explicit self-esteem and self-enhancement points to an underlying motivational process. Individuals generally strive for a positive self-view (Tajfel & Turner, 1997; Turner, Brown, & Tajfel, 1979). The more this is chronically available as in individuals with high self-esteem, the lower the motivation to self-enhance in a given situation. However, due to the correlational nature of these findings and because we did not find a similar effect in Study 1, these findings should be interpreted with caution.

In this paper we present two novel methods to measure self-enhancement that involve neither introspection nor any external standard of comparison. The method presented in Study 1 measures self-enhancement in a theory-driven way, whereas the method presented in Study 2 measures self-enhancement in a fully data-driven way (both on the individual or group-level) and is therefore informative about the exact dimension of enhancement. The traditional reverse correlation technique operates on 2D face images. Therefore, a specific random noise pattern affects different individuals' faces in different ways. A dark pixel interacts with the underlying face differently depending on the characteristics of that face. Therefore, this method would only allow extraction of individual self-recognition images. Here, we brought this technique to a new level by integrating it into the face space approach (Walker & Vetter, 2009, 2016). This approach works with 3D head data and thus allows application of the same random noise patterns or

vectors to different faces affecting these faces similarly. Therefore, we cannot only extract individual self-recognition images, but aggregate and compare them across different groups (here: males vs. females). Moreover, this novel method has the advantage that the extracted vector can be associated with meaning, in that we compare it with meaningful dimensions in our face space (here: agency and communion). One interesting endeavor for future research would, for example, be to compare self-enhancement effects across different cultures. Previous findings show enhancement effects only for Western but not for Asian individuals (Heine & Hamamura, 2007). Given that personality ascriptions from faces are highly cross-culturally shared and the method to manipulate perceived personality in faces has been successfully applied to faces from different cultural backgrounds (Walker et al., 2011), this method to very subtly measure self-enhancement allows self-perception vectors for individuals from different cultures to be developed individually and compared with each other as well as with various meaningful dimensions in the face space to investigate whether individuals from different cultures possibly enhance differently both regarding direction and degree of self-enhancement.

Conclusion

Taken together, the studies presented here add to the literature of self-enhancement by showing that participants do not only self-enhance regarding attractiveness (Epley & Whitchurch, 2008; Hancock & Toma, 2009), but also regarding a general valence dimension (i.e., likeability) and the Big Two personality dimensions, agency and communion. Moreover, they add to the literature of face processing by showing that the two basic dimensions of social perception (i.e., trustworthiness and dominance in the model of Oosterhof & Todorov, 2008 or the semantically and statistically similar dimensions of communion and agency in the Big Two personality model of Wiggins, 1991 used here) are not only fundamental when it comes to the

evaluation of unknown others based on their appearance, but also when it comes to the perception and recognition of individuals' own faces.

Importantly, these studies present two novel methods to implicitly and objectively measure the direction and degree of self-enhancement regarding various dimensions. These methods employ neither self-evaluations nor evaluations of or by others. By disposing of these explicit evaluations, these measures circumvent the problems of introspection, response bias, and the question of which of the two evaluations is more accurate than the other. Moreover, these methods are so subtle that they do not reveal the dimensions under investigation and, therefore, they are not susceptible for correction strategies. The reverse correlation method does not even require that the researcher knows the dimension of self-enhancement a priori but allows for explorative a posteriori analyses.

Therefore, we believe that the methods presented here can advance theory regarding self-enhancement in the long run, because they allow investigation of self-enhancement (and self-protection) regarding various dimensions (e.g., facial, personality, status, typicality of a certain group membership), various groups of individuals (e.g., from different cultural backgrounds, age groups), and individuals in different situations (e.g., by temporarily manipulating self-esteem or group membership), thus providing information about inter-group, inter-individual, and intra-individual differences in self-enhancement. With regard to the benefits of self-enhancement, such as being happy and caring about the self and others (Taylor & Brown, 1988), detecting the groups or individuals who are successful, and the situations that facilitate doing so, seems to be a critical endeavor for future research.

References

Abele, A. E. (2003). The dynamics of masculine-agentic and feminine-communal traits: Findings from a prospective study. *Journal of Personality and Social Psychology*, *85*(4), 768–776. http://doi.org/http://dx.doi.org/10.1037/0022-3514.85.4.768

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, *93*(5), 751–763. http://doi.org/http://dx.doi.org/10.1037/0022-3514.93.5.751

Abele, A. E., & Wojciszke, B. (2014). Communal and agentic content in social cognition: A dual perspective model. *Advances in Experimental Social Psychology*, *50*, 195–255. http://doi.org/10.1016/B978-0-12-800284-1.00004-7

Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, *55*(3), 387–395. http://doi.org/10.1037/0022-3514.55.3.387

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology*, *49*(6), 1621–1630. http://doi.org/10.1037/0022-3514.49.6.1621

Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. *The Self in Social Judgment*, *1*, 85–106.

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology*, *20*(1), 1–48. http://doi.org/10.1080/10463280802613866

Allison, S. T., Messick, D. M., & Goethals, G. R. (1989). On being better but not smarter than others: The Muhammad Ali effect. *Social Cognition*, *7*(3), 275–295.

http://doi.org/10.1521/soco.1989.7.3.275

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. http://doi.org/10.18637/jss.v067.i01

Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology*, *38*(6), 977–997. http://doi.org/10.1002/ejsp.487

Brown, J. D. (2012). Understanding the better than average effect: Motives (still) matter. *Personality and Social Psychology Bulletin*, *38*(2), 209–219. http://doi.org/10.1177/0146167211432763

Brown-Iannuzzi, J. L., Dotsch, R., Cooley, E., & Payne, B. K. (2017). The relationship between mental representations of welfare recipients and attitudes toward welfare. *Psychological Science*, *28*(1), 92–103. http://doi.org/10.1177/0956797616674999

Dalbert, C. (1992). Subjektives Wohlbefinden junger Erwachsener: Theoretische und empirische Analysen der Struktur und Stabilität. [Young adults' subjective well-being: Theoretical and empirical analyses of its structure and stability.]. *Zeitschrift Für Differentielle Und Diagnostische Psychologie*, *13*(4), 207–220.

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science*, *3*(5), 562–571. http://doi.org/10.1177/1948550611430272

Dotsch, R., Wigboldus, D. H. J., Langner, O., & van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, *19*(10), 978–980. http://doi.org/10.1111/j.1467-9280.2008.02186.x

Epley, N., & Whitchurch, E. (2008). Mirror, mirror on the wall: Enhancement in self-

recognition. *Personality and Social Psychology Bulletin*, *34*(9), 1159–1170. http://doi.org/10.1177/0146167208318601

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. http://doi.org/10.1037//0033-295X.102.4.652

Hancock, J. T., & Toma, C. L. (2009). Putting your best face forward: The accuracy of online dating photographs. *Journal of Communication*, *59*(2), 367–386. http://doi.org/10.1111/j.1460-2466.2009.01420.x

Heck, P. R., & Krueger, J. I. (2015). Self-enhancement diminished. *Journal of Experimental Psychology: General*, *144*(5), 1003–1020. http://doi.org/10.1037/xge0000105

Heine, S. J., & Hamamura, T. (2007). In search of east Asian self-enhancement. *Personality and Social Psychology Review*, *11*(1), 4–27. http://doi.org/10.1177/1088868306294587

Imhoff, R., Woelki, J., Hanke, S., & Dotsch, R. (2013). Warmth and competence in your face! Visual encoding of stereotype content. *Frontiers in Psychology*, *4*(386), 1–8. http://doi.org/10.3389/fpsyg.2013.00386

Janich, H., & Boll, T. (1982). *Übersetzung des Self-Esteem-Fragebogens von Rosenberg (1965)*. *Unpublished manuscript*, Fachbereich I - Psychologie, Universität Trier, Trier, Deutschland.

John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology*, *66*(1), 206–219. http://doi.org/10.1037/0022-3514.66.1.206

Keller, M. D., Reutner, L., & Walker, M. (2017, July). *Reverse correlation 2.0 – Combining a face space approach with up-to-date computer graphics.* Poster presented at the 18th meeting of the European Association of Social Psychology (EASP), Granada, Spain.

Kitayama, S., & Karasawa, M. (1997). Implicit self-esteem in Japan: Name letters and birthday

numbers. *Personality and Social Psychology Bulletin*, *23*(7), 736–742. http://doi.org/10.1177/0146167297237006

Kontsevich, L. L., & Tyler, C. W. (2004). What makes Mona Lisa smile? *Vision Research*, *44*(13), 1493–1498. http://doi.org/10.1016/j.visres.2003.11.027

Krueger, J., Ham, J. J., & Linford, K. M. (1996). Perceptions of behavioral consistency: Are people aware of the actor-observer effect? *Psychological Science, 7*(5), 259-264. http://doi.org/10.1111/j.1467-9280.1996.tb00371.x

Krueger, J. I., Heck, P. R., & Asendorpf, J. B. (2017). Self-enhancement: Conceptualization and assessment. *Collabra: Psychology*, *3*(1), 1-11. http://doi.org/10.1525/collabra.91

Krueger, J. I., & Wright, J. C. (2011). Measurement of self-enhancement (and self-protection). In M. D. Alicke & C. Sedikides (Eds.), *Handbook of Self-Enhancement and Self-Protection* (pp. 472-494). New York, NY: Guilford Press

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science*, *1*(2), 115–121.

Lewinsohn, P. M., Mischel, W., Chaplin, W., & Barton, R. (1980). Social competence and depression: The role of illusory self-perceptions. *Journal of Abnormal Psychology*, *89*(2), 203–212. http://doi.org/10.1037/0021-843X.89.2.203

Liu, J., Harris, A., & Kanwisher, N. (2002). Stages of processing in face perception: An MEG study. *Nature Neuroscience*, *5*(9), 910–916. http://doi.org/10.1038/nn909

Malle, B. F., & Pearce, G. E. (2001). Attention to behavioral events during interaction: Two actor-observer gaps and three attempts to close them. *Journal of Personality and Social Psychology. 81*(2), 278-294. http://doi.org/10.1037/0022-3514.81.2.278

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the

information employed for face classifications. *Cognitive Science*, *28*(2), 209–226. http://doi.org/10.1016/j.cogsci.2003.11.004

Mita, T. H., Dermer, M., & Knight, J. (1977). Reversed facial images and the mere-exposure hypothesis. *Journal of Personality and Social Psychology*, *35*(8), 597–601. http://doi.org/10.1037/0022-3514.35.8.597

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092. http://doi.org/10.1073/pnas.0805664105

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609. http://doi.org/10.1037/0022-3514.46.3.598

Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in the interplay of self-deceptive styles with basic traits and motives. *Journal of Personality*, *66*(6), 1025-1060. https://doi.org/10.1111/1467-6494.00041

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*. http://doi.org/10.1109/AVSS.2009.58

Perret, D. I., Lee, K. J., Penton-Voak, I., Rowland, D., Yoshikawa, S., Burt, D. M., … Akamatsu, S. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature*, *394*, 884–887.

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rhodes, G. (1986). Memory for lateral asymmetries in well-known faces: Evidence for configural information in memory representations of faces. *Memory & Cognition*, *14*(3),

209–219. http://doi.org/10.3758/BF03197695

Singh, S. P. (2014). Not all election winners are equal: Satisfaction with democracy and the nature of the vote. *European Journal of Political Research*, *53*(2), 308–327. http://doi.org/10.1111/1475-6765.12028

Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences*, *115*(37), 9210–9215. http://doi.org/10.1073/pnas.1807222115

Suitner, C., & Maass, A. (2008). The role of valence in the perception of agency and communion. *European Journal of Social Psychology*, *38*(7), 1073–1082. http://doi.org/10.1002/ejsp.525

Tajfel, H., & Turner, J. C. (1997). An integrative theory of intergroup conflict. In W. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Brooks/Cole: Pacific Grove.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193–210. http://doi.org/10.1037/0033-2909.103.2.193

Turner, J. C., Brown, R. J., & Tajfel, H. (1979). Social comparison and group interest in ingroup favouritism. *European Journal of Social Psychology*, *9*(2), 187–204. http://doi.org/10.1002/ejsp.2420090207

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, *43*(2), 161–204. http://doi.org/http://dx.doi.org/10.1080/14640749108400966

Walker, M., Jiang, F., Vetter, T., & Sczesny, S. (2011). Universals and cultural differences in

forming personality trait judgments from faces. *Social Psychological and Personality Science*, *2*(6), 609–617. http://doi.org/10.1177/1948550611402519

Walker, M., Schönborn, S., Greifeneder, R., & Vetter, T. (2018). The Basel Face Database: A validated set of photographs reflecting systematic differences in Big Two and Big Five personality dimensions. *PLoS ONE*, *13*(3), e0193190. http://doi.org/10.1371/journal.pone.0193190

Walker, M., & Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision*, *9*(11), 1–13. http://doi.org/10.1167/9.11.12

Walker, M., & Vetter, T. (2016). Changing the personality of a face: Perceived Big Two and Big Five personality factors modeled in real photographs. *Journal of Personality and Social Psychology*, *110*(4), 609–624. http://doi.org/10.1037/pspp0000064

Wiggins, J. S. (1991). Agency and communion as conceptual coordinates for the understanding and measurement of interpersonal behavior. In D. Cicchetti & W.M.Grove (Eds.), *Thinking Clearly about Psychology: Essays in honor of Paul E. Meehl, Vol. 1. Matters of public interest; Vol. 2. Personality ans psychopathology* (pp. 89–113). Minneapolis, MN, US: University of Minnesota Press.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598. http://doi.org/10.1111/j.1467-9280.2006.01750.x

Table 1

Reaction times in milliseconds (means, standard deviations) and linear and quadratic trend analyses (F, p, and η^2) testing how quickly participants recognize positive, original, and negative versions of their own faces among 8 distractor faces (reaction time task) in Study 1. The trend analyses were performed after logarithmic transformations of the mean values.

	M_N *(SD)*	M_{orig} *(SD)*	M_P *(SD)*	F_{linear}	p_{linear}	$\eta_p^2{}_{linear}$	$F_{quadratic}$	$p_{quadratic}$	$\eta_p^2{}_{quadratic}$
Attractiveness	1307.52* (470.03)	1167.86 (356.99)	1195.53 (423.53)	10.92	.002	.150	9.93	.003	.138
Likeability	1239.88* (381.81)	1167.86 (356.99)	1163.96 (362.36)	7.98	.006	.114	2.74	.103	.042
Agency	1307.98* (456.15)	1167.86 (356.99)	1164.94* (380.27)	14.51	<.001	.190	7.24	.009	.105
Communion	1186.75 (339.82)	1167.86 (356.99)	1216.48 (355.46)	0.99	.324	.016	2.96	.091	.046

Note. * indicate that single comparisons revealed significant differences from the original face at $\alpha = .05$.

Table 2

Distortions in self-recognition reflected in explicit choice (means, standard deviations) in Study 1. Values above 5 show distortions towards the positive pole, whereas values below 5 show distortions towards the negate pole of each dimension (multiples task).

	M (SD)	*t*	*p*	*d*
Attractiveness	5.67 (2.44)	2.20	.031	0.27
Likeability	6.02 (2.06)	3.95	.000	0.49
Agency	5.70 (2.38)	2.36	.021	0.30
Communion	5.53 (2.43)	1.75	.085	0.22

Table 3

Fixed effects for the mixed effects general linear model with the strength of the correlations between the random vectors and the agency [communion] vector as predictor and choice of agentic [communal] version as criterion in Study 2.

	Beta Estimate	*Std. Error*	*Z*	*p*
Agency				
Intercept	-.246	.047	-5.21	<.001
Correlation between agency and random vectors	1.210	.079	15.36	<.001
Communion				
Intercept	-.066	.045	-1.48	.138
Correlation between communion and random vectors	.706	.077	9.18	<.001

Note. Choice was treated as binomial, participants and random vectors were treated as random effects.

Table 4

Effects for the linear model with self-esteem as predictor and the correlation between individual self-enhancement vectors and agency [communion] as criterion in Study 2.

	Beta Estimate	*Std. Error*	*t*	*p*
Agency				
Intercept	.326	.141	2.31	.023
Self-esteem	-.083	.034	-2.42	.017
Communion				
Intercept	.321	.144	2.23	.028
Self-esteem	-.080	.035	-2.29	.024

A B

Figure 1. Two potential exemplar trials of the multiples self-recognition task in Study 1. The nine versions of the portrait slightly differ regarding perceptions of A) agency and B) communion. For reasons of visualization the three portraits in the first line depict the faces with highest levels of agency [communion] (left), the faces with lowest levels of agency [communion] (middle) and the faces with the second lowest levels of agency [communion] (right). The original face is presented in the middle row on the right.

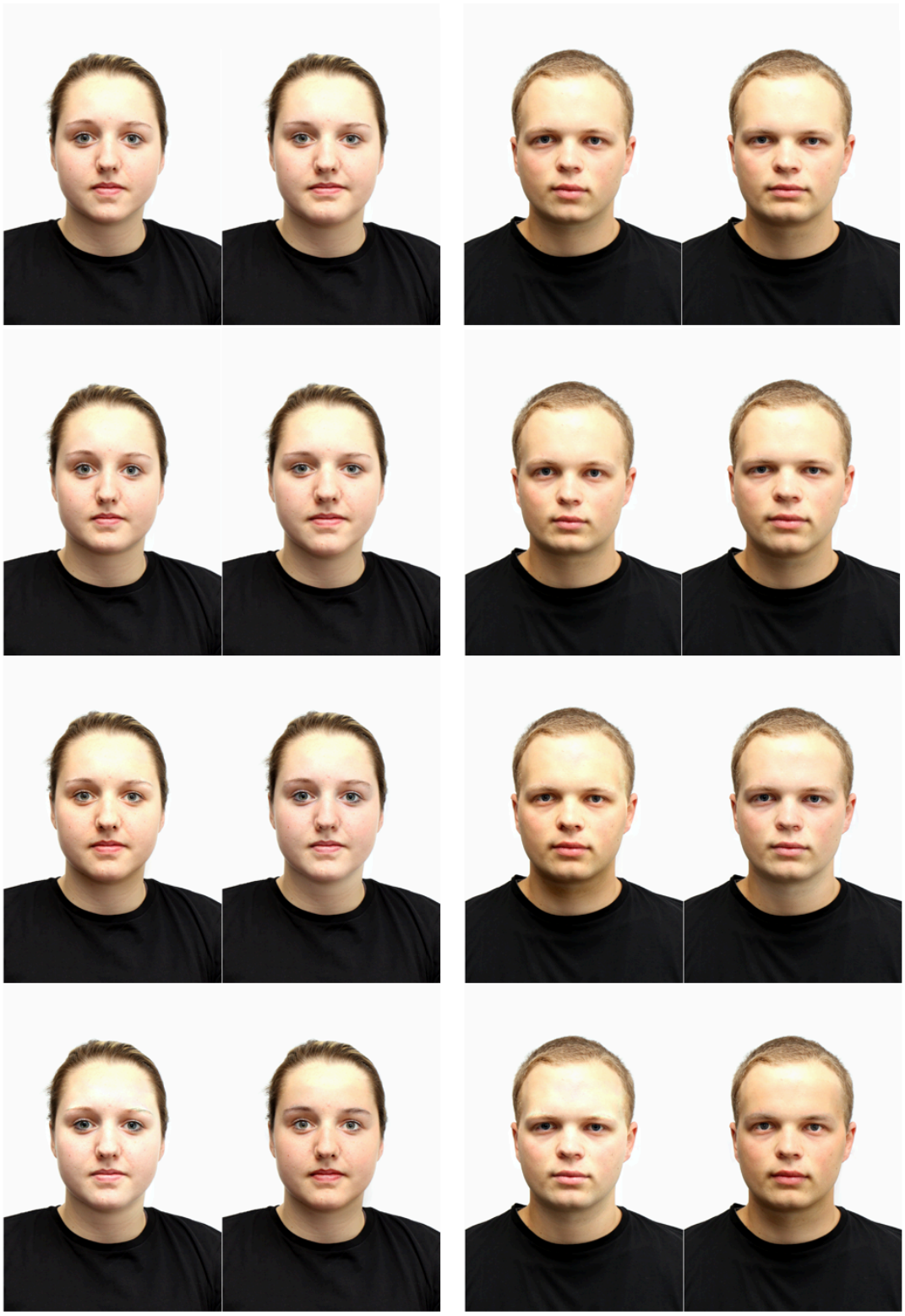


Figure 2. Exemplar trials of the image classification task in Study 2 for one female and one male participant. The first two rows represent two shape trials, the second two represent color trials.

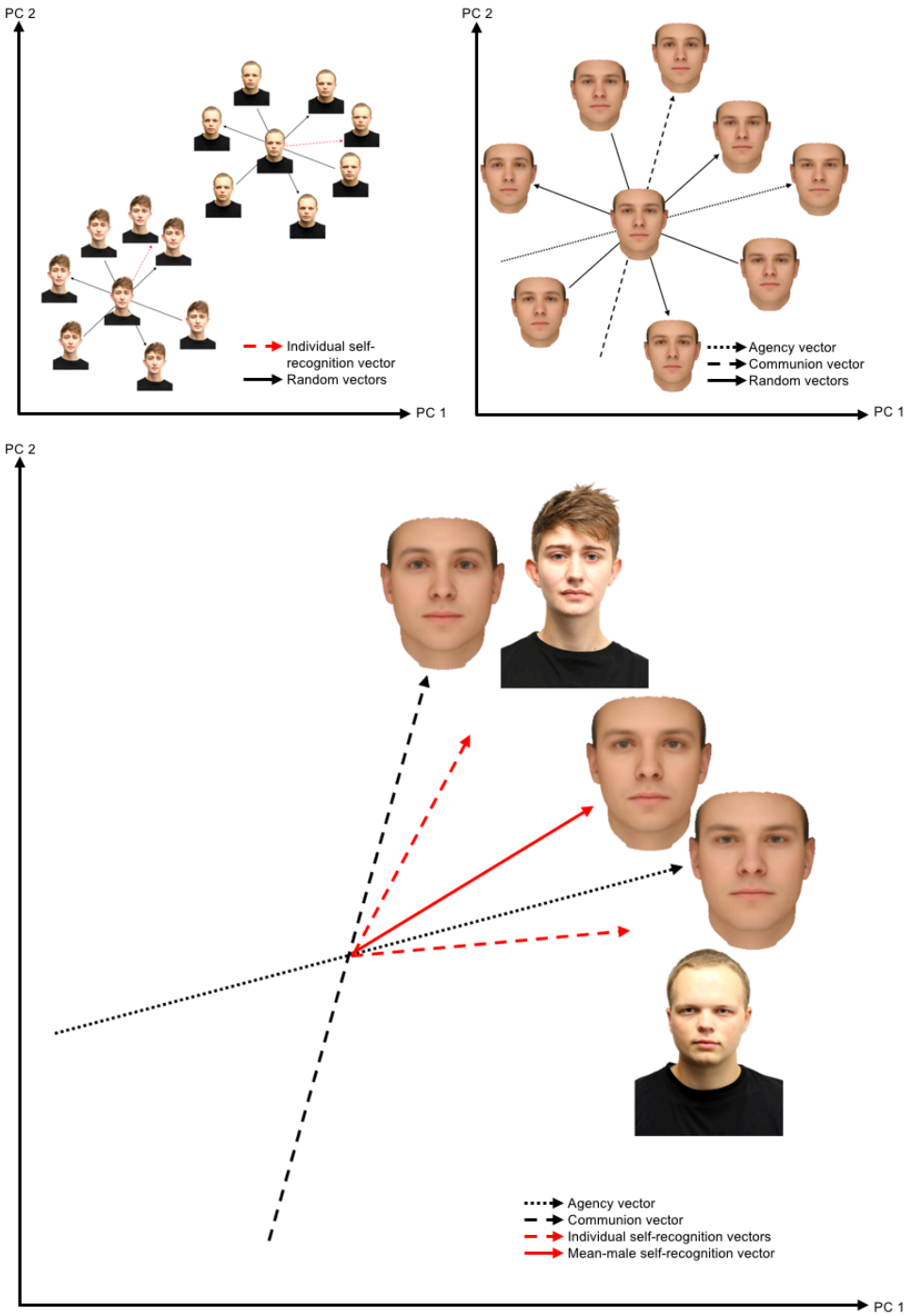


Figure 3. Visualization of fictitious individual self-recognition vectors (top left) and agency and communion vectors (top right) extracted from participants choice of randomly varying faces. Localization of fictitious individual self-recognition, agency, and communion vectors in the same face space allows determining on which dimensions individuals self-enhance (bottom).

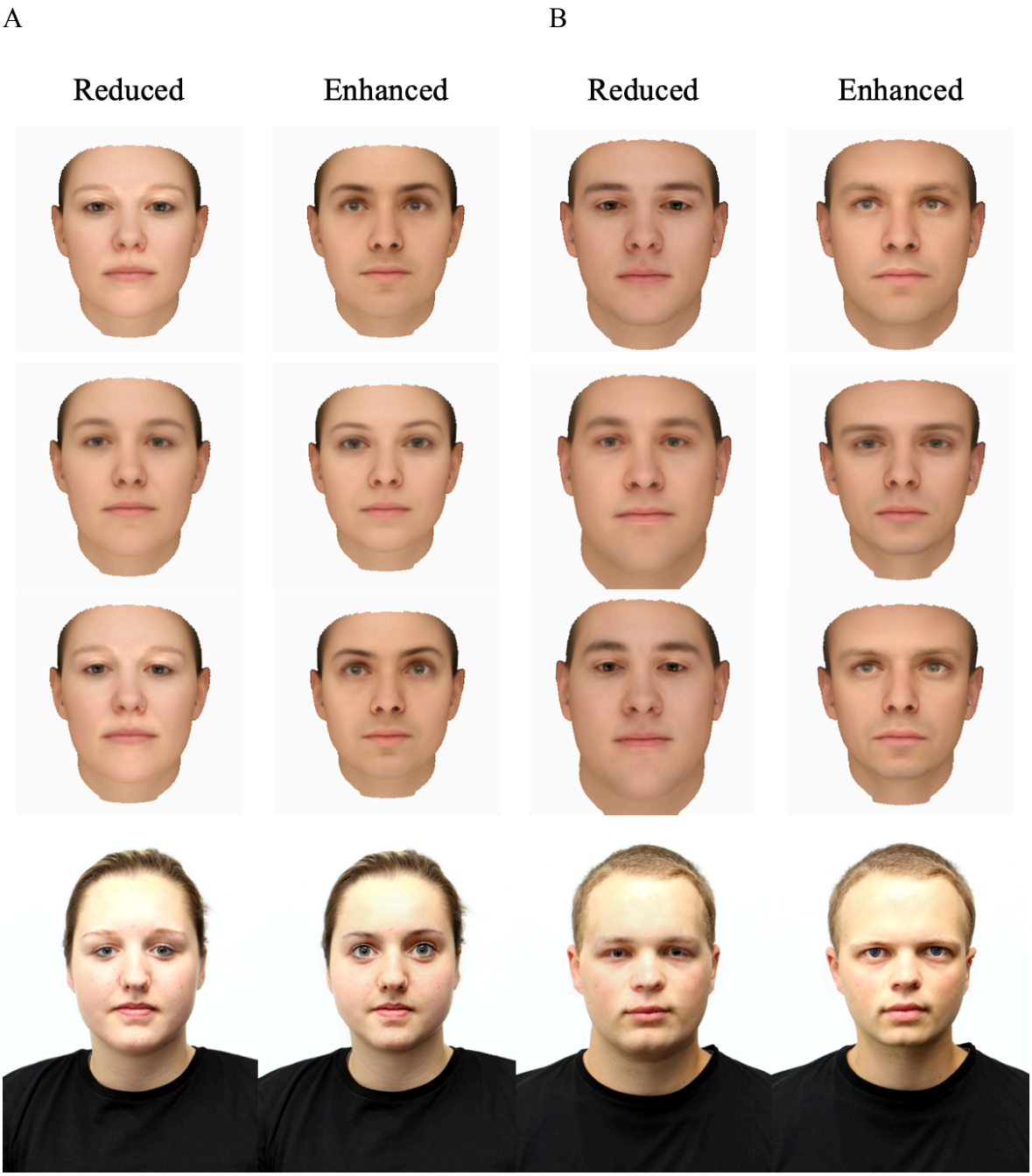


Figure 4. Visualization of the female (A) and male (B) color (row 1), shape (row 2), and full self-enhancement vector applied to the female (A) and the male (B) average face from the Basel Face Model (row 3) and to individual participants' faces (row 4).

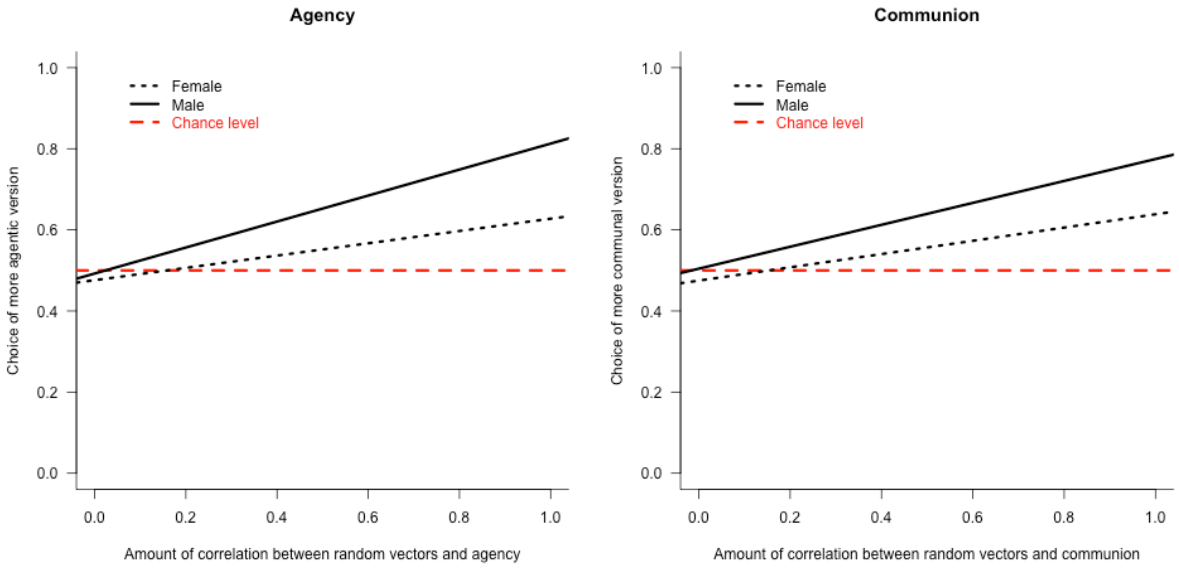


Figure 5. Enhancement tendencies on the agency and communion dimensions separately for male and female participants.

Appendix D

Stolier, R. M., Hehman, E., **Keller, M. D.**, Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 9210-9215. doi:10.1073/pnas.1807222115

This is a preprint of this manuscript, currently in press at the Proceedings of the National Academy of Sciences of the United States of America, version dated July 30th, 2018 (date will updated with preprint, old versions documented by OSF). Project data and analysis materials are available online: https://osf.io/z23kf/.

The conceptual structure of face impressions

Ryan M. Stolier[a], Eric Hehman[b], Matthias D. Keller[c], Mirella Walker[c], & Jonathan B. Freeman[a,d]

[a]Department of Psychology, New York University, 6 Washington Place, New York, NY 10003
[b]Department of Pyschology, McGill University, 845 Sherbrooke Street, Montreal, QC, Canada H3A 0G4
[c]Department of Psychology, University of Basel, Missionsstrasse 64A, 4055 Basel, Switzerland
[d]Center for Neural Science, New York University, 6 Washington Place, New York, NY 10003

Corresponding author:
Ryan M. Stolier or Jonathan B. Freeman
Department of Psychology
New York University
6 Washington Place
New York, NY 10003
Telephone: 212.998.7825
Email: rystoli@nyu.edu or jon.freeman@nyu.edu

Major classification: Social Sciences
Minor classification: Psychological and Cognitive Sciences

Abstract: 244 Words
Article body: 6147 Words
Number of figures: 2 figures (2 color)

Abstract

Humans seamlessly infer the expanse of personality traits from others' facial appearance. These face impressions are highly intercorrelated, within a structure known as 'face trait space'. Research has extensively documented the facial features that underlie face impressions, thus outlining a bottom-up fixed architecture of face impressions, which cannot account for important ways impressions vary across perceivers. Classic theory in impression formation emphasized that perceivers use their lay conceptual beliefs about how personality traits correlate to form initial trait impressions, for instance, where trustworthiness of a target may inform impressions of their intelligence to the extent one believes the two traits are related. This considered, we explore the possibility this lay 'conceptual trait space' —how perceivers believe personality traits correlate in others— plays a role in face impressions, tethering face impressions to one another and thus shaping face trait space. In Study 1, we found conceptual and face trait space explain considerable variance in each other. Study 2 found that participants with stronger conceptual associations between two traits judged those traits more similarly in faces. Importantly, using a face image classification task, Study 3 found participants with stronger conceptual associations between two traits used more similar facial features to make those two face trait impressions. Together, these findings suggest lay beliefs of how personality traits correlate may underlie trait impressions, and thus face trait space. This implies face impressions are not only derived bottom-up from facial features, but are also shaped by our conceptual beliefs.

Keywords: *face perception*, *impression formation*, *implicit personality theory*, *conceptual knowledge*, *dimensional models, social cognition*

Significance Statement

Current theory of face-based trait impressions focuses on their foundation in facial morphology, from which emerges a correlation structure of face impressions due to shared feature dependence, ‘face trait space’. Here, we proposed that perceivers’ lay conceptual beliefs about how personality traits correlate structure their face impressions. We demonstrate that ‘conceptual trait space’ explains a substantial portion of variance in face trait space. Further, we find that perceivers who believe any set of personality traits (e.g., trustworthiness, intelligence) are more correlated in others use more similar facial features when making impressions of those traits. These findings suggest lay conceptual beliefs about personality play a crucial role in face-based trait impressions, and may underlie both their similarities and differences across perceivers.

\body

THE CONCEPTUAL STRUCTURE OF FACE IMPRESSIONS

Humans naturally infer a broad range of personality traits from a face (1). From trustworthiness to creativity, we develop reliable impressions of others within seconds of seeing their face (2, 3). These face impressions influence our social behavior in situations as meaningful as election outcomes (4) and criminal sentencing (5).

Extensive research has documented how individual trait impressions are derived from morphological features of a face, for instance, that we infer both trustworthiness and submissiveness from babyfacedness (6). Naturally following, a central feature of face impressions is their highly intercorrelated structure (i.e., 'face trait space'), in which each trait impression is correlated with many others (1). Thus, current perspectives explain face impressions as derived by specific facial features, and face trait space as emergent from the degree to which different trait impressions share a similar featural basis (e.g., kindness and submissiveness also relate to babyfacedness, and thus both correlate with trustworthiness; 1). While such approaches have been highly valuable, they have tended to focus on a fixed architecture underlying face trait space – comprised of either two (1) or three (7) core dimensions – that are commonly assumed to not change across perceivers.

In this article, we propose that face impressions, and thus their correlations (face trait space), are further structured by perceiver lay theories of others' personality. Specifically, we propose that face impressions (e.g., intelligence) are also derived from the perception of other traits in a face (e.g., trustworthiness), insofar as a perceiver believes those two traits tend to correlate in other people. For example, a perceiver who believes the concept of trustworthiness is more related to the concept of intelligence may see a trustworthy face as more intelligent.

Research has long demonstrated that people hold rich lay conceptual associations of how they believe personality traits correlate in the world (in this article referred to as 'conceptual trait space'; 8, 9, 10). A common conceptual trait space has echoed throughout data-driven social perception research, where it has been long noted that a similar structure emerges across impression domains (face impressions, familiar person knowledge, stereotype content; 1, 11, 12-15). Classic theory in person perception emphasized the role of this conceptual trait space in shaping initial impressions (i.e., lay, or implicit personality theories; 16). For example, in seminal research of these questions, Asch (17) noted of his findings, "If a man is intelligent, this has an effect on the way in which we perceive his playfulness, happiness, friendliness" (p. 264). Yet, to our knowledge, such insights have not been directly applied or tested in understanding trait impressions of faces (though overlap in conceptual and face trait spaces has been observed towards romantic partner preferences; 18). If a perceiver's conceptual associations in part help scaffold face trait space, this may further formal models of face impressions generally, and an important implication would be that face trait space is dynamic across perceivers rather than representing any single fixed architecture (9).

Across several studies, we describe evidence that perceivers' beliefs in trait associations, or conceptual trait space, relate to their impressions of faces and in turn the structure of faces' trait space. First, we demonstrate broadly that face trait space reflects conceptual trait space, finding substantial overlap between the two (Study 1). Second, we find that perceivers' unique conceptual trait associations are related to the correlations of their individual face impression judgments (Studies 2 and 3). Lastly, we find perceivers' conceptual associations are related to the featural face space that underlies their impressions, which manifests in how they subjectively

perceive individual traits in the first place (Study 3). For all studies, all data and code are publicly available via the Open Science Framework (https://osf.io/z23kf/)..

Results

Study 1

Given a relatively common conceptual (19) and face (3) trait space between perceivers, they should show substantial overlap with one another on average if perceiver lay theories of personality shape their face impressions. It is possible that face trait space and conceptual trait space would not match. For instance, one can imagine the belief that dominant people are intelligent, responsible, and outgoing, yet the facial cues that give rise to dominance impressions may not give rise to intelligence impressions (1). These spaces could organize themselves by any number of factors that could structure trait concepts (e.g., valence, such as in a halo effect'; 11). Therefore, it is important to directly assess the correspondence of conceptual and face trait spaces. We first sought to empirically measure conceptual trait space (of the 13 traits used to estimate seminal models of face trait space; 1), and assess whether face trait space reflects its structure. To do so, we used representational similarity analysis (20), a powerful technique to assess similarity in such multivariate spaces (8, 9). In this technique, each trait space is represented as a similarity (i.e., correlation) matrix (pair-wise relations of all traits to each other; Fig. 1a,b), and then flattened into a vector of the unique pairwise similarity values between each trait. Because traits *within* a single matrix were measured on the same scale, similarities within each matrix were calculated using the standard distance metric of Pearson correlation, specifically, the pairwise correlations of trait judgments made of faces (see Materials and Methods). But because raw values in different matrices have different meanings, we assessed the correspondence *between* separate matrices (e.g., conceptual and face trait space) using Spearman rank correlation, which uses rank order rather than raw values (i.e., Pearson correlations) to estimate relationships between distances in the two spaces (20).

We measured conceptual and face trait space in two separate samples of participants. Each trait space was measured within a set of 13 personality traits used in seminal work quantifying face trait space: 'aggressive', 'caring', 'confident', 'dominant', 'egotistic', 'emotionally-stable', 'intelligent', 'mean', 'responsible', 'sociable', 'trustworthy', 'unhappy', and 'weird' (1). Similarities in the conceptual trait matrix were calculated using a straightforward pair-wise similarity rating: the average degree to which participants believed each unique pair-wise combination of personality traits are interrelated in other people (n = 113; e.g., trustworthy-dominant pair: 'If someone is trustworthy, how likely are they to be dominant?'; Fig. 1b, top row; see Materials and Methods). To estimate face trait space, 90 different faces were rated by participants on each of the 13 trait stimuli (n = 415; each participant randomly assigned to one of the 13 trait stimuli; Fig. 1b, bottom row). Indeed, as hypothesized, the conceptual and face similarity matrices explained a substantial amount of variance in one another (Spearman $\rho(76)$ = .82, ρ^2= .67, $p < .0001$, 95% CI = [.74, .88]; Fig. 1c). These results suggest that, when any two traits (e.g., caring and intelligent) are deemed more correlated in others, judgments of those traits in other people's faces exhibit a corresponding similarity or dissimilarity.

We replicated this relationship between conceptual and face trait matrices with a different face trait space, using trait judgment data from the original research defining the face trait space (1). (Note that 'egotistic' was removed in this analysis, as it was not present in this specific dataset). Indeed, a near-identical significant correlation between the conceptual and face trait model replicated this finding (Spearman $\rho(64)$ = .84, ρ^2= .71, $p < .0001$, 95% CI = [.75, .90]). Together, these results provide evidence for a strong correspondence between conceptual trait space and face trait space, consistent with a long history of research suggesting this correspondence (13-15, 18).

Study 2

Study 1 provides evidence that face trait space shares considerable structure with conceptual trait space (13). However, if conceptual associations play a role in shaping face trait space, perceivers' own face trait space should reflect their personal beliefs in how traits are conceptually associated. Meaning that while conceptual and face trait spaces were estimated on average across subjects in Study 1, Study 2 accounted for between-subject differences in trait associations, assessing the relationship between perceivers' idiosyncratic conceptual and face trait spaces (n = 206). This question is an important step in addressing whether perceivers' own conceptual trait associations influence their face impressions. By current perspectives (1, 6), overgeneralized facial cues (e.g., resting smile resemblance of a face) activate specific trait concepts (e.g., trustworthiness) identically across perceivers, due to adaptive associations between traits and those overgeneralized cues, and regardless of perceivers' conceptual association between the cue-related trait impression (e.g., trustworthiness) and other trait impressions made from the same face (e.g., dominance, creativity). Such perspectives do not predict that face impressions would relate to individual differences in conceptual associations, whereas our account does indeed predict this.

Each participant was randomly assigned to one unique pair from a subset of the pairwise combinations in Study 1: 'assertive', 'caring', 'competent', 'creative', 'self-disciplined', and 'trustworthy'. (Due to practical limitations in measurement, note this looks through a pinhole at this process, only investigating single trait-pairs per subject, rather than measuring the entirety of their trait spaces need to acquire a full picture of this process). Participants evaluated faces on both assigned traits, then later provided a conceptual similarity judgment between those traits, as in Study 1. Thereby, in this study participants served as the unit of analysis, with a score for their

conceptual and face trait similarity. To test our hypothesis, we correlated participants' idiosyncratic face and conceptual trait similarities. Participants' conceptual similarity rating for a given trait pair was correlated with how similar those traits were judged in faces (Spearman $\rho(204) = .34$, $\rho^2 = .12$, $p < .0001$, 95% CI = [.21, .46]; Fig. 2a). These findings demonstrate a correspondence between how similar a participant idiosyncratically deems two traits and how similarly the participant judges those traits in others' faces. Thus, the results replicate and extend those of Study 1, documenting correspondence between conceptual and face trait spaces on an individual-level.

Study 3

We have seen that conceptual trait space and face trait space explain considerable variance in one another (Study 1), and further, explain individual differences in each other (Study 2). These findings have testable implications for face impressions. If two different trait impressions are more or less correlated with one another, the facial features that typically evoke those impressions are likely to shift towards or away from one another, fundamentally altering the featural space underlying face impressions. In other words, perceivers who differ in the degree of conceptual association between traits would "see" these traits differently in faces. For instance, someone who believes agreeable people are often open to experience may make both impressions from faces based on more similar visual features. Someone who does not think agreeable people are often open to experience, on the other hand, may make both impressions based on less similar features.

To test this possibility, we applied a recently advanced reverse-correlation technique, which allowed us to estimate the facial features underlying participants' perceptions of traits in a data-driven manner (21). Using this technique, we obtained a featural vector in face space that

represents each participant's visual representation of each trait. Thereby we estimated the perceived visual similarity of different traits in faces for each participant. Identical to Study 2, we then tested whether a participant's idiosyncratic conceptual similarity between any two traits related to the visual similarity in features that evoke those specific traits for the participant. Each participant (n = 185) was randomly assigned to one unique pair from the unique pairwise combinations of the big-five factor personality traits: 'agreeable', 'conscientious', 'extroverted', 'neurotic', and 'open to experience'. These traits were used to increase generalization of the findings of Studies 1-2, and also given prior success in deriving these traits within the statistical face model we used (21). Participants performed a forced choice image classification task (e.g., (22) for each trait assigned, then later provided their idiosyncratic conceptual similarity rating between those traits. Accordingly, our data included each participant as the unit of analysis as in Study 2, with a score for their conceptual and face trait similarity. Consistent with our hypotheses, a participant's conceptual similarity between two traits was correlated with the visual similarity in facial features associated with those traits (Spearman $\rho(183) = .40$, $\rho^2 = .16$, $p < .0001$, 95% CI = [.27, .51]; Fig. 2b). These findings show that the extent to which the visual features underlying each trait impression are more or less similar to those of other trait impressions relate to perceivers' own conceptual association between those traits. We illustrate this in Figure 2b, in which we present the 'agreeable' and 'open to experience' classification images produced from two individual participant responses. For example, a participant who deems agreeableness and openness to be more conceptually related tends to "see" these traits as visually more similar in people's faces (i.e., uses similar features to make these impressions; see Fig. 2b).

Discussion

Together, our findings suggest that perceiver lay theories of personality may play an important role in face-based trait impressions. First, we found that conceptual trait space and face trait space explain a considerable amount of variance in each other (Study 1). The relationship between conceptual trait associations and face trait associations is further evidenced by our findings that face impression judgments correlate within perceivers to the degree they believe those traits are more similar conceptually (Study 2). Lastly, we found that conceptual trait associations predict the visual features perceivers use to infer those traits in others' faces. Thus, our findings provide correlational evidence suggesting that face impressions (e.g., intelligence) are partly derived from one another (e.g., trustworthiness), to the extent perceivers believe those traits are correlated in other people.

The current results provide several important contributions to theories of face impressions. The role of conceptual trait associations in face impression processes adds a crucial top-down layer to what have been predominately feature-driven bottom-up models (1, 6). If face impressions are derived from one another by way of their conceptual associations, this process may explain considerable correspondence in the structure of face impressions across perceivers (Study 1; 1, 3), given similar correspondence in conceptual trait associations across perceivers (19). Above and beyond this commonality, this process may explain important individual differences in perceivers' face impressions and trait space (Studies 2 and 3), to the extent their conceptual trait associations vary. As such, the findings bolster recent proposals arguing that face trait space may reflect a dynamic integration of not only intrinsic facial-feature covariation but also conceptual associations, stereotypes, and other social cognitive factors (9). Interestingly, the notion that individual differences in conceptual associations between traits shapes perceptions

comports well with seminal person perception research that posited a role of 'implicit personality theory' in non-face trait impressions (16, 17). The results therefore suggest that these classic insights with respect to general impression-formation patterns (outside of face perception) may apply to face-based trait impressions as well.

A common correlated structure of trait impressions has been observed not only in face impressions, but also in person knowledge and group-level stereotypes (1, 11, 12). This structure extends further to explain mental state inferences (23), as well as neural representations during social perception (24). That perceptions across domains share such similar structure is striking, and perhaps telling of a common cognitive basis for correlated social perceptions (13-15). Future research could directly investigate the role of conceptual trait spaces in shaping the structure of person perception in other domains, such as abstract representations of others (e.g., outside the domain of face evaluation; 11) and social groups (12), including the possibility of empirically connecting these various spaces together. Understanding the contribution of perceiver conceptual trait associations to social perception across these domains may be paramount to understanding real world social behavior that is quite consequential. Dimensions of both face impressions and group stereotypes are highly consequential, in situations serious as such as election outcomes (4, 25) and criminal sentencing (5, 26). Future research should assess whether important individual and cultural differences in conceptual trait space alter critical social decisions. In the mean time, the use here of RSA (20) is noteworthy approach to assess similarities across these domains, where it has also benefited comparisons of conceptual trait spaces with domains distant as actual personality (27) and social categorization (28).

With respect to such dimensions, the results may provide a parsimonious explanation for cases in which their correlations may cease to be independent and shift. In one example, trait

impressions of less familiar others may be more intercorrelated and lower dimensional than those of familiar others (29, 30). It may be the case that perceivers rely more on their conceptual trait space, in which trait judgments are highly correlated, to make impressions of unfamiliar others when more specific person-knowledge is unavailable. For example, additional information about targets allows trait dimensions of sociability and morality, typically linked to one another (1, 12), to become orthogonal (31). This account could also generalize to explain models of trait impressions in intergroup contexts. For instance, use of a conceptual trait-space to make wide personality inferences towards unfamiliar outgroup members may underlie systematically biased (32-34) and therefore homogenous trait impressions (35). Yet, increased information about targets may disengage use of the conceptual trait space (i.e., individuation; 33). Another notable example is the more negative relationship between trustworthiness and dominance impressions of female compared to male faces (36), presumably due to stereotypes linking female likability with submissiveness (37). Our findings suggest that unique conceptual trait spaces, such as when considering different social groups (e.g., conceptual associations between traits when regarding females vs. males), may lead to differential associations between face impressions. Future research could measure shifts in conceptual trait space in different social contexts, to assess whether variations in face trait and group-level trait space emerge from a conceptual basis.

There are important limitations of the current work. Most notably, the correlational nature of our design precludes any strong inference about the causal impact of conceptual knowledge on face trait space. Alternative possibilities exist, including face impressions shaping conceptual trait space. At face value, it seems unlikely that individual differences in face impression correlations (due to mere featural processing of the same face stimuli) could exert such a consistent influence on participant conceptual associations between personality traits. This is

especially the case given perceivers would have to track whether impressions of faces from one task somehow reflected those in the second separate task, and there is a considerable lack of awareness concerning which features underlie perceivers' judgments (2, 38-40). Yet our current data cannot exclude these possibilities. Future research should seek causal evidence of conceptual knowledge's influence on face trait space by manipulating conceptual knowledge directly.

Another noteworthy limitation is the use of language, trait concept terms such as 'trustworthiness', to measure both face impressions and conceptual associations. This issue has been central to longstanding debates concerning the origins of lay personality theory models, in which researchers have debated whether measured trait concept associations are merely semantic in nature, rather than underlain by beliefs about actual personality traits of others (for a review, see (16). If perceivers' trait term semantic associations (e.g., believing the words 'kind' and 'sociable' mean the same thing) are all that is behind their conceptual and face trait associations, similarity in conceptual and face trait spaces may be an artifact of language and uninteresting for understanding social behavior. Speaking against this possibility, many researchers have found evidence that trait concept correlations are independent of semantic features, and argued semantic explanations do not obviate socially meaningful and consequential trait relations (41, 42). Nonetheless, such ruling out has not been applied in the current domain of face impressions, and future research should evaluate this concern in this context. Future research could examine measure whether the significance of a trait impression changes, for instance whether conceptual shifts in intelligence impressions impact its affective (e.g., evaluative priming) or behavioral (e.g., hiring decisions) consequences for perceivers.

In conclusion, we found that lay conceptions of personality traits are strongly related to trait impressions based on other people's facial appearance. The common structure that emerges across perceivers in face impressions (1, 3) has considerable resemblance to commonly shared conceptual trait structure (11). Beyond any such shared structure, individual differences in perceivers' conceptual trait associations are related to the unique structure of their face impressions and the features that underlie them. Together, these findings suggest the way we infer personality traits from faces are not only determined by the physical appearance of a face, but also by our own lay conceptual beliefs regarding the personality of others.

Materials and Methods

Data, analysis code, and results are all available and hosted by the Open Science Framework (https://osf.io/z23kf/). Data may be downloaded, and results reproduced via Jupyter notebooks available in the repository.

Study 1

Participants.

Face trait space. We collected face impression data from 415 subjects via Amazon Mechanical Turk (demographic data missing for 1 subject; all United States residents; all primary English-speakers; M_{age} = 34.23 years, SD_{age} = 12.27 years; 260 Female, 146 Male, 2 other, 5 decline; 316 White, 33 Black, 28 Asian, 38 other). Participants were randomly assigned to evaluate one personality trait in all face stimuli, and were therefore divided roughly equally between all 13 personality trait conditions (≈32 participants per trait condition). Subjects were financially compensated for their participation, and they gave informed consent. This experiment was approved by the University Committee on Activities Involving Human Subjects at New York University.

Conceptual trait space. We collected conceptual trait association data from 113 subjects via Amazon Mechanical Turk (demographic data missing for 1 subject; all United States residents; all primary English-speakers; M_{age} = 36.34 years, SD_{age} = 11.14 years; 72 Female, 40 Male; all White). Subjects were financially compensated for their participation, and they gave informed consent. This experiment was approved by the University Committee on Activities Involving Human Subjects at New York University.

Stimuli.

Face stimuli. All stimuli were taken from the Chicago Face Database (43). Face stimuli included 90 portrait photographs of young white male individuals with neutral facial expressions. These stimuli were also used in Study 2. A secondary analysis looked at a face trait similarity model derived from seminal work in face trait space measurement. In this study (1), 66 faces (female and male) from the Karolinska Directed Emotional Faces face database (44) were rated on each trait (besides 'egotistic'; 1 – 9 Likert-type scale; e.g., 1 – 'Not at all trustworthy', 9 – 'Extremely trustworthy'). See the original publication for additional details (data available upon request from the authors' web database; https://tlab.princeton.edu/databases/).

Personality trait stimuli. We chose 13 personality traits that independent groups of participants evaluated in faces and in conceptual similarity. These traits were those used in the seminal work assessing face trait space (1). In this work, these traits were chosen as those unique but also spontaneously elicited during face impressions (with the exception of 'dominance', which was included by the researchers). These traits included: 'aggressive', 'caring', 'confident', 'dominant', 'egotistic', 'emotionally-stable', 'intelligent', 'mean', 'responsible', 'sociable', 'trustworthy', 'unhappy', and 'weird'.

Protocol. See Supporting Information for detailed task instructions.

Face trait space task. Participants were informed they would partake in a study examining how people perceive others. Each participant was randomly assigned to evaluate only 1 of the 13 personality trait stimuli in faces. In the task, participants rated each of the 90 face stimuli on the personality trait they were assigned (1 – 7 Likert-type scale; e.g., 1 – 'Very untrustworthy', 4 – 'neutral', 7 – 'Very trustworthy'). Following the face trait rating task, participants completed a general demographics survey and completed the experiment.

Conceptual trait space task. Participants were informed they would partake in a study on how different personality traits correlate in the world. Participants evaluated the conceptual relationship of each trait-pair in the 13 trait stimuli (1 – 7 Likert-type scale, 1 – 'Not at all likely', 4 – 'Neutral', 7 – 'Very likely'), presented in both order given the wording of the item question (e.g., 'trustworthy – dominant' and 'dominant – trustworthy'). Therefore, there were a total of 156 trials for each participant ($P(13,2) = 156$). Following the face trait rating task, participants completed a general demographics survey and completed the experiment.

Data preparation and analysis. All analyses were conducted with scientific and statistical libraries in Python. No subjects were removed from these data before analysis. To assess whether face trait space reflects conceptual trait space, we applied a quantitative method from systems neuroscience, representational similarity analysis (RSA; 20). As a straightforward explanation, this analysis measured the correlation between trait-pair similarity matrices as measured in the face trait and conceptual trait tasks. An intuitive description of this process is to correlate the unique values of two different similarity matrices together, assessing the similarity between the two correlation matrices. Therefore we may assess whether the similarity of face trait judgments reflects the pattern of how similar those traits are conceptually conceived. See a detailed explanation of RSA in the Supporting Information.

Study 2

Participants. We collected face impression data from 206 subjects via Amazon Mechanical Turk (original $n = 213$; 2 subjects dropped due to task incompletion; 5 subjects dropped due to failure to follow task instructions; all United States residents; all primary English-speakers; $M_{age} = 29.78$ years, $SD_{age} = 6.81$ years; 102 Female, 65 Male, 1 decline; gender data from 38 participants missing due to a data collection error; 160 White, 17 Black, 9

Asian, 20 other). Subjects were financially compensated for their participation, and they gave informed consent. This experiment was approved by the University Committee on Activities Involving Human Subjects at New York University.

Stimuli.

Face stimuli. Face stimuli were identical to those collected in our data in Study 1 (see Study 1 methods).

Personality trait stimuli. We chose a diverse set of trait stimuli somewhat deviating from those in Study 1 to assess generalizability. Trait stimuli included: 'assertive', 'caring', 'competent', 'creative', 'self-disciplined', and 'trustworthy'. We used all pairwise combinations of these trait pairs (for a total of 15 unique possible trait-pairs). Participants were randomly assigned to one of the 15 total trait-pair combinations.

Protocol. Both face trait and conceptual trait tasks were largely identical in design within themselves to those in previous studies (see Study 1 methods). A major distinction is that in this study, each participant both provided face trait and conceptual trait data. Each participant was randomly assigned to one of 15 trait-pairs (the unique combinations of 6 trait stimuli: 'assertive', 'caring', 'competent', 'creative', 'self-disciplined', and 'trustworthy'). First, participants evaluated all face stimuli on both assigned traits. They evaluated all stimuli on one trait first, followed by the other. The order of which trait was first evaluated was randomly determined per subject. In total, participants therefore completed 180 trials of face impressions. From this data, we were able to measure the correlation of face impressions within each subject. Second, participants provided conceptual trait association ratings for their assigned trait-pair. As participants only evaluated the similarity of two traits to one another (as compared to the many trait-pairs in Study 1), there were only 2 trials in the conceptual trait task. Instructions and item

design were identical to those used in Study 1. Following these tasks, participants completed a general demographics survey.

Data preparation and analysis. In Study 2, we ask whether the amount to which each perceiver associates two trait concepts relates to the correlation between those trait impressions in faces. That is, we intended to test whether perceivers with weaker/stronger conceptual trait associations also show more weakly/strongly correlated face impressions. To do so, within each perceiver, we calculated two variables: their conceptual and face trait associations (see Supporting Information). To test our hypothesis, we calculated the Spearman correlation between participant face trait and conceptual trait associations (Spearman correlation used so as to not assume a strictly linear relationship between distances in the two spaces) (20). Analyses were conducted across trait-pair terms, to assess the tendency of conceptual trait associations to relate to face impression correlations, across trait-pairs in general.

Study 3

Participants. We collected face trait image classification data from 186 subjects via Amazon Mechanical Turk (original n = 194; 9 subjects removed due to task incompletion; all United States residents; all primary English-speakers; M_{age} = 33.89 years, age data for 1 subject missing, SD_{age} = 8.6 years; 113 Female, 72 Male, 1 decline; 139 White, 21 Black, 11 Asian, 15 other). Subjects were financially compensated for their participation, and they gave informed consent. This experiment was approved by the University Committee on Activities Involving Human Subjects at New York University.

Stimuli.

Face stimuli. First, we created an average face from 100 female and 100 male faces from the Basel Face Model (45). Within the shape and the color space spanned by these 200 faces, we

created 100 vectors randomly varying face shape and 100 vectors randomly varying face color. Separately applying these 200 vectors to the average face in both positive and negative direction resulted in 200 pairs of faces or 200 classification trials, respectively.

Personality trait stimuli. Personality trait stimuli included the big-five personality traits ('agreeable', 'conscientious', 'extroverted', 'neurotic', 'open to experience'), due to their successful use in prior work with this statistical face manipulation technique (21). Furthermore, these new trait stimuli allowed us to even further diversify our trait stimuli to strengthen inferences of generalizability. We used all pairwise combinations of these trait pairs (for a total of 10 unique possible trait-pairs). Participants were randomly assigned to one of the 10 total trait-pair combinations.

Protocol. The overall structure of the study was similar to the structure used in Study 2. Each participant both provided face trait and conceptual trait data. Participants were randomly assigned to one of the 10 trait-pair permutations (i.e., one of the pairwise combination of the Big Five traits, varying in order by which trait was listed first to counterbalance the task below). Each participant completed four image classification tasks. They first performed a shape and a color task for the first trait they were assigned to, followed by a shape and a color task for the second trait they were assigned to. All four tasks comprised 100 trials. In each trial, participants were presented with two faces horizontally adjacent to one another on the same page (i.e., random vector applied to the average face in positive direction and in negative direction), and asked to indicate which of the two faces looks more extreme regarding the trait in question (e.g., which face looks more 'agreeable'). Following the image classification task, participants provided conceptual trait association ratings for trait-pairs assigned. This task was identical to that in Study 2. Lastly, participants completed a general demographics survey.

Data preparation and analysis. In Study 3, we ask whether the amount to which each perceiver associates two trait concepts is related to the correlation between those traits' face space feature vectors (i.e., 'face trait vectors') estimated from the image classification task. That is, we tested whether perceivers with weaker/stronger conceptual trait associations actually see traits less/more similarly in faces. Within each perceiver, we calculated two variables: their face trait vectors' correlation and conceptual trait associations (see Supporting Information). To test our hypothesis, we calculated the Spearman correlation between participant face trait vectors and conceptual trait associations (Spearman correlation used so as to not assume a strictly linear relationship between distances in the two spaces; 20). Analyses were conducted across trait-pair terms, to assess the tendency of conceptual trait associations to predict face trait vector correlations, across trait-pairs in general.

Figure Captions

Figure 1. *Comparison of conceptual and face trait spaces*. In Study 1, we quantitatively assess the correspondence in structure of conceptual and face trait space. Panel a provides an illustration of conceptual (top row) and face trait space models (bottom row) with multidimensional scaling. In our analysis, we test correspondence of each trait space by the Spearman correlation of unique values above the diagonal of their similarity matrices (panel b; conceptual, i.e., 'How likely is a person with one trait likely to have the other'; and face, i.e., how correlated are face impressions of one trait with another). Analyses indicated the trait spaces overlap in structure substantially (13), Spearman $\rho(76) = .82$, $p < .0001$. Although the analysis was carried out using Spearman correlation, for illustrative purposes only Pearson correlation is depicted. MDS plots are organized by k-means clustering within each trait space, whereas both similarity matrices are sorted by the k-means clustering solution of the conceptual matrix for comparability.

Figure 2. *Conceptual trait associations relate to visual similarity in facial features used for trait impressions*. If lay conceptual beliefs about how personality traits correlate shape face impressions, perceivers' who believe two traits are more related (e.g., 'agreeableness' related to 'openness') should infer a trait from a face (e.g., 'agreeableness') to the extent they infer the other trait simultaneously from that face (e.g., 'openness'), and thus see those traits more similarly in faces (e.g., illustration in panel b, right). In Study 2, we found participants who believed two personality traits were more correlated in others (e.g., 'agreeable people are often open') also judged faces along those two traits more similarly (e.g., judged faces they perceived agreeable to also be open), Spearman $\rho(204) = .34$, $p < .0001$ (panel a). In Study 3, participants with stronger conceptual associations between two traits (e.g., 'agreeable people are often open')

also used similar facial features to make those trait impressions of faces (e.g., facial features underlying agreeableness impressions were more similar to those underlying openness impressions; measured via image classification task), Spearman $\rho(183) = .40$, $p < .0001$ (panel b, left). Although the analysis was carried out using Spearman correlation, for illustrative purposes only Pearson correlation is depicted. In panel b (right), we also present two example participants to illustrate these findings, where a participant with high conceptual associations between agreeableness and openness (top row) sees those traits in faces more similarly than a participant low in that association (bottom row).

Acknowledgments

We thank John Andrew Chwe and Clodagh Cogley for assistance in materials development and data collection, and Andreas Morel-Forster and Thomas Vetter for providing assistance and materials regarding the Basel Face Model. This work was supported in part by National Institutes of Health fellowship grant F31-MH114505 (R. M. S.) and National Science Foundation research grant BCS-1654731 (J. B. F.).

References

1. Oosterhof NN & Todorov A (2008) The functional basis of face evaluation. *Proceedings of the National Academy of Sciences* 105:11087-11092.
2. Todorov A, Pakrashi M, & Oosterhof NN (2009) Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition* 27(6):813-833.
3. Hehman E, Sutherland CA, Flake JK, & Slepian ML (2017) The Unique Contributions of Perceiver and Target Characteristics in Person Perception.
4. Todorov A, Mandisodza AN, Goren A, & Hall CC (2005) Inferences of competence from faces predict election outcomes. *Science* 308(5728):1623-1626.
5. Wilson JP & Rule NO (2015) Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological science* 26(8):1325-1331.
6. Zebrowitz LA & Montepare JM (2008) Social Psychological Face Perception: Why Appearance Matters. *Soc Personal Psychol Compass* 2(3):1497.
7. Vernon RJ, Sutherland CA, Young AW, & Hartley T (2014) Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences* 111(32):E3353-E3361.
8. Tamir DI & Thornton MA (2018) Modeling the Predictive Social Mind. *Trends in cognitive sciences* 22(3):201-212.
9. Stolier RM, Hehman E, & Freeman JB (2018) A Dynamic Structure of Social Trait Space. *Trends in cognitive sciences* 22(3):197-200.
10. Osgood CE (1952) The nature and measurement of meaning. *Psychological bulletin* 49(3):197.

11. Rosenberg S, Nelson C, & Vivekananthan P (1968) A multidimensional approach to the structure of personality impressions. *Journal of personality and social psychology* 9(4):283.

12. Fiske ST, Cuddy AJ, Glick P, & Xu J (2002) A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology* 82(6):878-902.

13. Todorov A, Said CP, Engel AD, & Oosterhof NN (2008) Understanding evaluation of faces on social dimensions. *Trends in cognitive sciences*.12(12):pp.

14. Fiske ST, Cuddy AJ, & Glick P (2007) Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences* 11(2):77-83.

15. Oldmeadow JA, Sutherland CA, & Young AW (2013) Facial stereotype visualization through image averaging. *Social Psychological and Personality Science* 4(5):615-623.

16. Schneider DJ (1973) Implicit personality theory: A review. *Psychological bulletin* 79(5):294.

17. Asch SE (1946) Forming impressions of personality. *The Journal of Abnormal and Social Psychology* 41(3):258.

18. South Palomares JK, Sutherland CA, & Young AW (2017) Facial first impressions and partner preference models: Comparable or distinct underlying structures? *British Journal of Psychology*.

19. Kuusinen J (1969) Factorial invariance of personality ratings. *Scandinavian journal of psychology* 10(1):33-44.

20. Kriegeskorte N, Mur M, & Bandettini P (2008) Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience* 2:4.

21. Walker M & Vetter T (2016) Changing the personality of a face: Perceived Big Two and Big Five personality factors modeled in real photographs. *Journal of personality and social psychology* 110(4):609.

22. Dotsch R, Wigboldus DH, Langner O, & van Knippenberg A (2008) Ethnic out-group faces are biased in the prejudiced mind. *Psychological science* 19(10):978-980.

23. Gray HM, Gray K, & Wegner DM (2007) Dimensions of mind perception. *Science* 315(5812):619-619.

24. Tamir DI, Thornton MA, Contreras JM, & Mitchell JP (2016) Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences* 113(1):194-199.

25. Hehman E, Carpinella CM, Johnson KL, Leitner JB, & Freeman JB (2014) Early processing of gendered facial cues predicts the electoral success of female politicians. *Social Psychological and Personality Science* 5(7):815-824.

26. Johnson SL, Eberhardt JL, Davies PG, & Purdie-Vaughns VJ (2006) Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychological science* 17:383-386.

27. Lay CH & Jackson DN (1969) Analysis of the generality of trait-inferential relationships. *Journal of personality and social psychology* 12(1):12.

28. Stolier RM & Freeman JB (2016) Neural pattern similarity reveals the inherent intersection of social categories. *Nature neuroscience* 19(6):795-797.

29. Thornton MA & Mitchell JP (2017) Theories of Person Perception Predict Patterns of Neural Activity During Mentalizing. *Cerebral cortex*:1-16.

30. Koltuv BB (1962) Some characteristics of intrajudge trait intercorrelations. *Psychological Monographs: General and Applied* 76(33):1.

31. Brambilla M, Rusconi P, Sacchi S, & Cherubini P (2011) Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology* 41(2):135-143.

32. Kunda Z & Thagard P (1996) Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological review* 103:284-308.

33. Fiske ST & Neuberg SL (1990) A continuum model of impression formation from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology* 23:1–74.

34. Brewer MB (1988) A dual process model of impression formation. *A Dual-Process Model of Impression Formation: Advances in Social Cognition*, eds Srull TK & Wyer RS (Erlbaum, Hillsdale, NJ), Vol 1, pp 1-36.

35. Quattrone GA & Jones EE (1980) The perception of variability within in-groups and out-groups: Implications for the law of small numbers. *Journal of personality and social psychology* 38(1):141.

36. Sutherland CA, Young AW, Mootz CA, & Oldmeadow JA (2015) Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated. *British Journal of Psychology* 106(2):186-208.

37. Glick P & Fiske ST (1996) The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of personality and social psychology* 70(3):491-512.

38. Freeman JB, Stolier RM, Ingbretsen ZA, & Hehman EA (2014) Amygdala responsivity to high-level social information from unseen faces. *The Journal of Neuroscience* 34(32):10573-10581.

39. Ambady N, Bernieri FJ, & Richeson JA (2000) Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in experimental social psychology, Vol 32*, (Academic Press, San Diego, CA), pp 201-271.

40. Tskhay KO & Rule NO (2013) Accuracy in categorizing perceptually ambiguous groups: A review and meta-analysis. *Personality and Social Psychology Review* 17(1):72-86.

41. Borkenau P (1992) Implicit Personality Theory and the Five-Factor Model. *Journal of Personality* 60(2):295-327.

42. Block J, Weiss DS, & Thorne A (1979) How relevant is a semantic similarity interpretation of personality ratings?

43. Ma DS, Correll J, & Wittenbrink B (2015) The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* 47(4):1122-1135.

44. Lundqvist D, Flykt A, & Öhman A (1998) The Karolinska directed emotional faces (KDEF). *D ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*.

45. Paysan P, Knothe R, Amberg B, Romdhani S, & Vetter T (2009) A 3D face model for pose and illumination invariant face recognition. *Advanced video and signal based surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, (Ieee), pp 296-301.

Figures

Figure 1

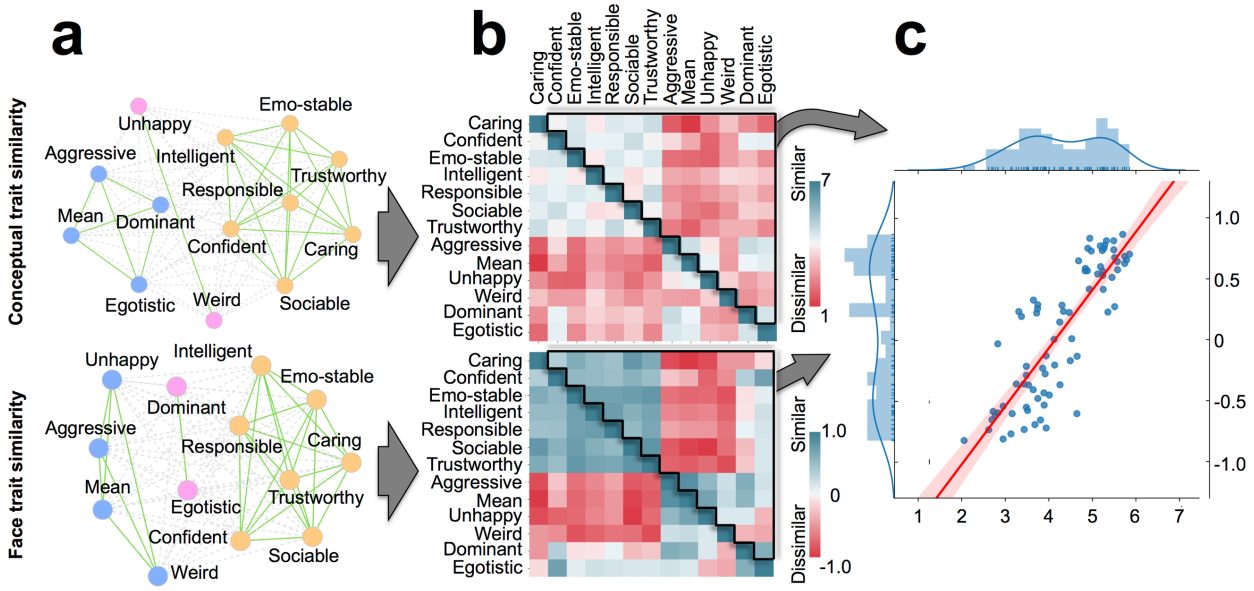
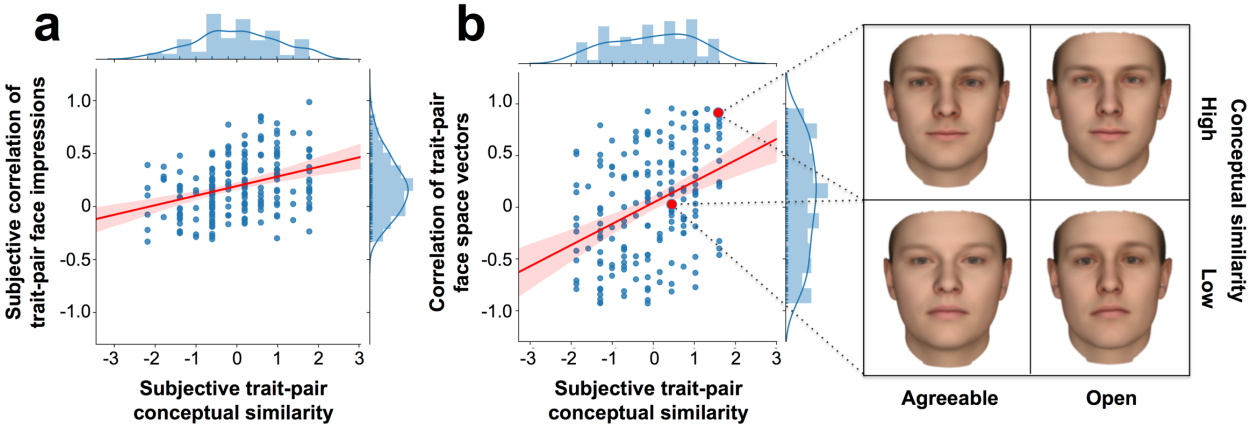


Figure 2



Supplementary Information

Study 1 Methods

Face trait space task protocol. Our specific instructions to participants were, "In this task, we ask you to indicate how [TRAIT STIMULUS] a number of different people look. You will see a person's face, and are asked to judge their likely personality traits merely from their face. Importantly, go with your gut feeling. We all make snap judgments of others constantly, so feel free to report what you think about the person based on their face. Please respond quickly with your gut feeling. There are no right or wrong answers."

Conceptual trait space task protocol. Our specific instructions to participants were, "In the following task, you will be presented with a series of adjective pairs. These are human personality traits. You will be asked to rate the likelihood that individuals with one of the traits possess the other trait." After several clarifications and examples of the task, participants began the task. Each trial item asked, "Given that an individual possesses one trait, how likely is it that they possess the other?", then presented the two trait stimuli for that trial separated by a hyphen (e.g., 'trustworthy – dominant').

Data preparation and analysis. In Study 1, to performed representational similarity analysis, we created a similarity matrix for each of our models – one for face trait space, one for conceptual trait space. Here we outline specific calculations underlying these matrices, which are also visible and reproducible in analysis scripts on the manuscript OSF page. To create our face trait similarity model (i.e., matrix), we calculated the average of each trait rating for each of the 90 face stimuli (leaving us with 13 trait ratings per each of 90 face stimuli). Then, we calculated the Pearson correlation between each vector of face ratings per trait condition, giving us the correlation (i.e., similarity) between each trait-pair in face trait ratings (Fig. 1a,b). Next, we created the conceptual trait similarity model (i.e., matrix). The pairwise similarity between each trait pair was simply calculated as the average rating of each unique trait-pair combination within and across subjects (i.e., average rating of participant belief that traits are likely shared in people; e.g., average of 'trustworthy – dominant' and 'dominant – trustworthy' within and across subjects). From this we create a similarity model between all trait-pairs as measured conceptually (Fig. 1a,b). To perform our analysis, we correlate the face trait and conceptual trait similarity models with one another. First, we obtain the unique similarity values from the diagonal of the similarity matrices (omitting redundant values from the symmetrical matrices, as well as the diagonal, in which each trait is always perfectly similar to itself). This creates a vector of similarity values per model. Next, we perform a Spearman rank correlation between the two models (as this is robust to similarity measurement idiosyncrasies across measurement modalities, e.g., face evaluations and conceptual trait ratings). (Figure 1 provides a conceptual illustration of this; for more detailed discussion and example of this analysis strategy in the context of face trait space).

Study 2 Methods

Data preparation and analysis. In Study 2, we estimated face and conceptual trait associations per participant. Here we outline specific calculations underlying these matrices, which are also visible and reproducible in analysis scripts on the manuscript

OSF page. To estimate their face trait association, we calculated the Pearson correlation coefficient between both trait evaluations of the face stimuli within each participant (between the vectors of their impressions of all face stimuli one each of the two traits they were assigned). To estimate their conceptual trait associations, we averaged the two conceptual trait items. Therefore a single dataset was created including data from participants across all trait-pair combinations.

Study 3 Methods

Data preparation and analysis. In Study 3, per participant we calculate their face trait vectors' correlation, and conceptual trait associations. Here we outline specific calculations underlying these matrices, which are also visible and reproducible in analysis scripts on the manuscript OSF page. To estimate their face trait vectors' correlation, we first calculated for each participant the two face trait vectors (per trait assigned to a participant) resulting from the four image classification tasks (each face trait vector combining information from the shape and color task per trait). To review, in each trial participants were presented with two faces: the same single average base face (which is represented as a vector of facial feature values), one adding and one subtracting the same random manipulation to its facial features (by applying a random noise facial feature vector to that of the base face, thus changing the appearance of the face in two directions along a random set of features in each trial). To calculate each trait vector, we averaged across the noise feature vectors (across 100 shape and 100 color vectors) that corresponded to the faces each participant selected. This provided a face trait vector per each trait assigned to a participant, comprised of the values for each feature participants had been tracking as belonging to the trait they sought to classify in the task. Finally, as a measure of similarity between individuals' face trait vectors, we calculated the Pearson correlation coefficient between the two extracted vectors. Thus, this correlation value is a measure of the similarity in facial features participants used to classify each trait, where a higher value signifies the participant used similar features to identify each trait. To estimate their conceptual trait associations, we averaged the two conceptual trait items. Therefore, a single dataset was created including data from participants across all trait-pair combinations.

Appendix E

Curriculum Vitae

Matthias David Keller

Education

09/ 2013 – 07/2015	***Master of Science in Psychology, University of Basel, CH*** Major: Social, economic and decision psychology
09/ 2011 – 07/2013	*Bachelor of Science in Psychology, University of Basel, CH*
09/ 2007 – 07/2011	*Bachelor of Science in Sport and History, University of Basel, CH*

Professional career

Since 09/2015	***Scientific Assistant & PhD Candidate*** Department of Social Psychology, University of Basel, CH
12/2012 – 08/2015	*Research Assistant* Department of Social Psychology, University of Basel, CH
07/2014 – 06/2015	*Research Assistant* Department of Social Psychology, University of Bern, CH
07/2012 – 08/2012	*Research Intern* Department of Social Psychology, University of Basel, CH

Publications

Content of dissertation:

Keller, M. D., Reutner, L., Greifeneder, R., & Walker, M. (2019). *Faces evoking emotions stereotypically triggered by groups: Developing an advanced reverse correlation technique*. Manuscript under review.

Stolier, R. M., Hehman, E., **Keller, M. D.**, Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences, 115*, 9210-9215. doi:10.1073/pnas.1807222115

Walker, M. & **Keller, M. D.** (2019). Beyond attractiveness: A multimethod approach to study enhancement in self-recognition on the Big Two personality dimensions. *Journal of Personality and Social Psychology*. Advance online publication. doi:10.1037/pspa0000157

Rudert, S.C., **Keller, M. D.**, Hales, A. H., Walker, M., & Greifeneder, R. (2019). *Who gets ostracized? A personality perspective on risk and protective factors of ostracism*. Manuscript in revision.

Additional research output:

Keller, M. D., Reutner, L., Walker, M., & Greifeneder, R. (2019). *Sexualized but not objectified – When do women react negatively towards sexualized advertisements*. Manuscript in revision.

Jaffé, M.*, **Keller, M. D.***, Jeitziner, L.*, & Walker, M., (2019). *The differences in faces do make a difference: Perceptions and preferences of diversity in ascribed personality traits*. Manuscript under review.
*shared first author

Walker, M. & **Keller, M. D.** (2019). *Moral character matters. How act and actor (facial) characteristics impact moral judgements*. Manuscript under review.