Journal of Experimental Social Psychology 100 (2022) 104277

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Differences in faces do make a difference: Diversity perceptions and preferences in faces

Mariela E. Jaffé [1,*], Loris Jeitziner [2], Matthias D. Keller [3], Mirella Walker [4]

Center for Social Psychology, University of Basel, Switzerland

ARTICLE INFO

Editor: Jack Rachael

Keywords:
Objective diversity
Perceived diversity
People perception
Facial information
Diversity beliefs

ABSTRACT

Throughout previous research focusing on individuals' diversity perception, it remains somewhat unclear which attributes (i.e., objective diversity) are reflected in perceptions of diversity. This manuscript investigates whether individuals consider objective differences in ambiguous facial information (which are not related to gender or race) when making diversity judgments and decisions. Throughout seven studies, facial information of group members was manipulated to appear more similar or different in regards to personality and information unrelated to Big 5 dimensions, while race, gender, and age were kept constant. Study 1a provides support that objective differences in facial information related to perceived personality traits is validly reflected in perceptions of diversity. Study 1b shows that results regarding the Big 5 can be replicated in an ensemble-coding setup. Studies 2a and 2b replicate this result, additionally showing that objective differences in facial information unrelated to the Big 5 are reflected in perceptions of diversity, too. Focusing on perceived extraversion, Study 3 reveals that individuals select faces differing (similar) in extraversion information in order to assemble a diverse (homogeneous) team. Study 4 investigates diversity choices in an ambiguous setting, showing that individuals who more strongly believe in the value of diversity are more likely to assemble a team that is objectively diverse regarding facial information. Study 5 indicates that the association between diversity in facial information and choices deteriorates if other attributes such as gender are varied too. The impact of the results for research is highlighted and discussed.

Diversity research focuses on differences regarding attributes on which individuals can be perceived to differ (Williams & O'Reilly, 1998). Classic examples of such attributes are gender, race, age, educational background, or socioeconomic status. When looking at groups, however, members will most likely differ not only on one but on several attributes and will be similar on others. Attributes may include the classically studied demographic attributes, but also personality, values, knowledge, and skills. When asking a team member how diverse they perceive their team to be (Shemla & Meyer, 2012) or an independent observer to judge the perceived diversity of the team, how do they make their judgment? Rephrased more concretely: When making diversity judgments, which attributes do people take into account? This question is important as it is perceived and not objective diversity that people will most likely act upon (Hobman, Bordia, & Gallois, 2003) and which can impact support for national and organizational policies related to diversity (Daniels, Neale, & Greer, 2017).

In this manuscript we aim to provide some answers to this question by investigating perceptions of diversity of groups. Previous research has studied the impact of demographic attributes such as gender, race, or age (Alt, Goodale, Lick, & Johnson, 2019; Daniels et al., 2017; Harrison, Price, Gavin, & Florey, 2002; Phillips, Slepian, & Hughes, 2018) on perceived group diversity. In this manuscript, however, we go beyond demographic variables that are strongly associated with diversity and are eventually presented as categories (as compared to continuous variables) by focusing on variance in facial information. We test whether ambiguous differences in group members' faces may impact diversity perceptions, too. Other variables within each group, such as gender, age, and race, are kept constant and the changes in facial

* Corresponding author at: Center for Social Psychology, University of Basel, Missionsstrasse 64a, 4055 Basel, Switzerland.
E-mail address: mariela.jaffe@unibas.ch (M.E. Jaffé).
[1] University of Basel
[2] Fachhochschule Nordwestschweiz (FHNW)
[3] LINK
[4] Pädagogische Hochschule Luzern

https://doi.org/10.1016/j.jesp.2021.104277
Received 17 May 2021; Received in revised form 29 November 2021; Accepted 24 December 2021
Available online 18 January 2022

information are not indicative for these categories.

As we vary facial information only, we provide ambiguous information that is unrelated to identity categories and that needs to be interpreted by participants. Therefore, it may or may not be included in their perceived diversity judgments. Next to studying the impact of facial information on diversity perceptions, we test whether differences in facial information can be considered when making diversity choices and whether individuals' diversity beliefs predict preferences for facially more diverse groups. Results from this research allow us to broaden our understanding of potential drivers of perceived diversity, which is likely the reality that people act upon in groups in their everyday lives (e.g., Hobman et al., 2003).

1. A brief overview of perceived diversity

Perceptions of diversity are presumably shaped by an interaction between objective differences regarding different attributes and the cultural or individual construal of diversity (Shemla & Meyer, 2012). When investigating perceived diversity, different approaches are possible. Cunningham (2007) conducted research where he specifically asked participants (sport coaches) about perceived differences between themselves and other coaches in regards to age and race. Here, correlations between objective and perceived demographic differences were positive and significantly different from zero (Cunningham, 2007). Harrison et al. (2002) investigated the association between objective and perceived diversity in regards to so called surface- and deep-level diversity variables. Surface-level diversity is defined in their work as differences in overt demographics, such as age, gender, or race (Harrison, Price, & Bell, 1998). Deep-level diversity is defined as differences among team members' psychological characteristics, including personalities, values, and attitudes (Harrison et al., 1998). When investigating the different concepts and their association with perceived diversity in business students doing team work, the researchers found that objective surface-level diversity influenced perceived surface-level and deep-level diversity influences perceived deep-level diversity (Harrison et al., 2002). However, when taking a closer look, the correlation tables reveal that differences in age, gender, and race as well as in marital status were clearly associated with perceived surface-level diversity. Yet, when focusing on perceived deep-level diversity the association is less clear. More specifically, only differences in the evaluation of task meaningfulness (the personal salience and importance of a team's project) and outcome importance (the value of getting a good grade for the team members) were associated with individuals' deep-level diversity perceptions, whereas differences in regards to conscientiousness and values were not significantly related (Harrison et al., 2002). One explanation could be that, as Harrison and Klein (2007) argued, individuals may simply lack the necessary information to accurately assess diversity, especially when considering attributes that are not easily visible.

Both studies investigated perceived diversity when looking at coaches judging themselves in relation to their peers (Cunningham, 2007) and students judging their student work groups (Harrison et al., 2002). In both cases, however, the teams might have differed in more variables than those studied and the participants making the judgments also had further knowledge and experiences from previous interactions, which might have impacted their judgments. Different research therefore turned to a more controlled setting, where participants were presented with pictures of faces of team members and asked for diversity judgments (e.g., Alt et al., 2019; Daniels et al., 2017; Haberman, Lee, & Whitney, 2015; Phillips et al., 2018).

Phillips et al. (2018), for example, showed participants faces of team members that varied in race, gender, or dominance and tested whether differences in facial properties would impact diversity judgments. Participants first saw a target group, then a comparison group, and were then asked to judge whether the latter was more or less diverse than the former. Participants were able to perform correct and above chance level diversity judgments for all three dimensions (see Phillips et al., 2018). The results highlight people's ability to perform ensemble-coding, meaning that they can extract group information (such as variance or diversity) for groups of faces.

This work was conceptually replicated: Alt et al. (2019) focused on gender and again presented faces of different numbers of women and men in a group. Participants' gender diversity judgments correctly reflected the ratio of men and women, and the authors further showed that a higher men to women ratio was associated with higher perceptions of threat (Alt et al., 2019). Daniels et al. (2017), furthermore, investigated perceptions of diversity based on gender and age information in face images, again showing that higher objective diversity was associated with higher perceived diversity, but also indicating spillover effects, such as that more diversity in regards to race resulted in the perception of more gender diversity, too (see Experiment 1; Daniels et al., 2017).

Variables such as race, gender, and age (that are all closely related to definitions and understanding of diversity) seem to impact perceptions of diversity. The relation between variables such as values or personality and perceptions of diversity seems to be less clear (see Harrison et al., 2002). Would differences in personality or general facial features also contribute to perceived diversity? One line of research that speaks to this question is work on the perception of emotions. The research highlights that people can make correct inferences about the variance in emotions depicted in a group of faces (Haberman et al., 2015). This provides a promising basis indicating that non-demographic variables can also be used to form variance judgments and eventually also diversity judgments (see also Phillips et al., 2018). Different emotions, however, may be easily interpretable. Could similar findings also occur in regards to values or personality, or even differences in faces, that are not easily interpretable?

One reason that makes the study of such effects difficult, is that, unlike gender, race, and emotions, variables such as personality are not as easily detectable from visual cues such as pictures (Harrison & Klein, 2007). Interestingly, people still have the impression that they can make judgments about somebody's values or personality from a picture (even when seeing the picture of a face for only 100 milliseconds; see Todorov, Pakrashi, & Oosterhof, 2009; Willis & Todorov, 2016). It therefore remains an open question whether, if differences in demographic attributes such as gender and race are kept constant, more ambiguous facial cues could impact diversity perceptions of groups. This manuscript aims to find answers to this open question, allowing for a broader outlook on which information is relevant for judgments regarding perceived diversity.

2. Personality judgments based on individuals' faces

One area that has studied the impact of cues on subsequent perceptions or judgments is the research on face perception. When individuals do not have access to background information, they may rely on other people's faces to make a judgment, for example, in regards to somebody's personality (Freeman & Ambady, 2011; Kubota & Ito, 2007; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). Although research has shown that the validity of such personality ascriptions is at the very most negligible (Olivola & Todorov, 2010), people spontaneously form relatively stable impressions (Klapper, Dotsch, van Rooij, & Wigboldus, 2016) and there is a high consensus among perceivers regarding these ascriptions (Oosterhof & Todorov, 2008), even across different cultures (Walker, Jiang, Vetter, & Sczesny, 2011). Using this consensus in ascriptions of Big Five personality dimensions from faces, statistical face models have been developed to systematically isolate and manipulate those characteristics in faces that impact, for example, the degree of extraversion perceived from a face (Walker & Vetter, 2016).

Applying these models to photographs of real faces systematically manipulates facial information and results in natural-looking portraits that are perceived as more or less extraverted, more or less conscientious, more or less open, more or less neurotic, and more or less agreeable (Walker, Schönborn, Greifeneder, & Vetter, 2018). Exemplary portrait pictures where facial information is varied can be found here: https://bfd.unibas.ch/en/det/validation-data/ (see also Basel Face

Database; Walker et al., 2018). This method allows the objective facial information of a person to be changed (e.g., the color, size, shape, localization of facial components), without changing the person itself, meaning that gender, age, and race and other identity-related information are kept constant. It is further worth mentioning that these manipulations are very subtle and that it is hard to name the objective facial information that is changing and that individuals use when making their judgments (see Walker & Vetter, 2016).

Interestingly, so far, the abovementioned research has focused on the manipulation and subsequent perceptions of *singular faces* and not on perceptions of *sets of faces* combined to a *group*. While research on "face perception has clearly demonstrated that visual processes play an essential role in how people think about other individuals, a corresponding revolution has yet to occur in research on groups" (Phillips, Weisbuch, & Ambady, 2014, p. 102; and see Alt & Phillips, 2021, for an overview on people perception). At the same time, the same authors highlight that "group perceptions, whether in ancient caves or in contemporary board rooms, should enable people to succeed in organizations and social life more generally" (Phillips et al., 2014, p. 102). Individuals therefore should be able to form group perceptions, eventually also regarding diversity (see Alt et al., 2019; Alt & Phillips, 2021; Daniels et al., 2017; Haberman et al., 2015; Phillips et al., 2018).

The face modelling methods mentioned above (e.g., Walker et al., 2018) provide the ideal setting to test our research question, namely whether objective differences in facial information that are not related to race or gender influence diversity perceptions of groups. Facial information can be manipulated on the individual picture level. Two groups can then be created out of the same individuals' pictures, which are then either manipulated to all look more extraverted (resulting in objectively higher levels of homogeneity in facial information) or differ in regards to extraversion by manipulating half of the pictures to appear more extraverted and the other half of pictures to appear more introverted (resulting in objectively higher levels of diversity in faces). If diversity perceptions are influenced by these objective facial information differences and not only by classic demographic attributes, the latter (compared to the former) group should be perceived as more diverse.

3. Overview of hypotheses and studies

In the research presented in this manuscript we used face modelling to create group pictures that are objectively more or less diverse in regards to facial information of the individual group members, while keeping gender and race constant. The manipulations used to change facial features were created to make a portrait appear more or less extraverted, conscientious, neurotic, agreeable, or open to experience. Although the changes affect perceived personality (and not actual personality), we always refer to modelling as changing the objective features and personality traits of the group. Based on the individual portrait pictures, we create more or less diverse groups and study the impact of objective differences in faces on perceptions and decisions. Within this research, we test three hypotheses:

First, we aimed to test whether diversity perceptions reflect stronger versus weaker objective differences in facial information of group members when demographic attributes such as gender, race, and age are kept constant (Hypothesis 1). We started the project with a stronger focus on objective differences in facial information that referred to differences such as Big 5 perceived personality traits (Hypothesis 1*, tested in Study 1a). Results from our later work (Studies 2a and 2b), however, indicated that this focus could be broadened to differences in facial information more generally that impact perceptions of diversity. This is why we broadened Hypothesis 1* to Hypothesis 1, and focus on differences in facial information in general and not on facial information related to Big 5 personality traits only. We predict that objective diversity (compared to homogeneity) in ambiguous facial information results in higher levels of perceived diversity (focus of Studies 1a, 1b, 2a, and 2b).

Second, moving from perceptions to behavior, we investigated whether individuals actively use differences in facial information when asked to assemble a more diverse versus a more homogeneous team (Hypothesis 2). We predict that individuals instructed to assemble more diverse (compared to homogeneous) teams are more likely to choose groups that are more different (compared to similar) in facial information. Study 3 was designed to test Hypothesis 2.

Third, we were interested in individuals' choice behavior when no instructions were provided. How likely is it that individuals will spontaneously assemble an objectively more versus less diverse team? In this context we assumed that diversity beliefs are an important predictor. We hypothesized that individuals who more strongly (compared to less strongly) believe in the value of diversity assemble more (compared to less) diverse teams. We tested Hypothesis 3 in Study 4.

Additionally, we tested whether facial information would still be used when being asked to assemble a diverse team when other, more obvious attributes such as gender varied, too. We did not have a specific hypothesis regarding this question, but aimed to exploratorily test how the association between facial variation and diversity choices may differ when variables such as gender are not held constant as in Studies 3 and 4. Study 5 was designed to investigate this question.

Seven studies were designed to test these hypotheses. While Study 1a was not preregistered, all of the predictions, conditions, dependent variables, and analyses for Studies 1b, 2a, 2b, 3, 4, and 5 have been preregistered at aspredicted.org. All studies have been conducted online. This entails that the screen size and the visual angle were not controlled, which might contribute to error variance in our studies. We disclose all measures, manipulations, and exclusions and indicate the method of determining the final sample size. Furthermore, all datasets are available and will be shared upon request.

4. Study 1a

Study 1a was designed to test whether individuals' perceptions of diversity are informed by differences in facial information that are related to perceived personality (Hypothesis 1*). To test this assumption, individuals were presented with group pictures that consisted of four individuals each. Each group picture was either manipulated to appear more diverse or less diverse in regard to their members' perceived personality (see Walker et al., 2018). Participants were asked to estimate how similar or different they perceived the group members to be. These evaluations were then used to analyze whether ambiguous objective differences in facial information would be associated with perceptions of diversity when looking at and judging the group pictures.

4.1. Method

4.1.1. Participants

Study 1a was conducted as an online survey and advertised as a "group impressions" study on Prolific (Prolific Academic, 2018). Our sample consisted of 110 participants. Based on predefined exclusion criteria, we excluded 2 participants because they indicated reasons to not use their data, 5 participants because they rated themselves lower than 5 on our carefulness scale (1 = *not carefully at all*, 7 = *very carefully*) and another 3 participants because they reported display errors when looking at the presented group pictures. That resulted in a final sample of 100 participants (58 male, 42 female; $M_{age} = 32.22$, $SD_{age} = 11.99$).

As Study 1a was more exploratory and we did not have access to specific effect size estimates informing our research question, we refrained from calculating an a priori power analysis. However, according to a sensitivity power analysis (calculations made using G*Power; Faul, Erdfelder, Lang, & Buchner, 2007), a minimal effect size of $d = 0.28$ could be detected under standard criteria ($\alpha = 0.05$, $1 - \beta = 0.80$, two-tailed) with our sample of 100 participants.

4.1.2. Design

Study 1a builds on a within design with three factors: diversity of

group (manipulated as diverse vs. homogeneous, meaning less diverse), personality factor on which diversity was manipulated (Big Five; see John & Srivastava, 1999; McCrae & Costa, 1997; Walker et al., 2018) and group number (three different group constellations were used: Group 1, Group 2, Group 3). The resulting design is a 2 (diversity) x 5 (personality dimension) within subjects design, with group number as a control factor. Perceived diversity served as dependent variable.

4.1.3. Materials

To manipulate diversity on facial information we used 12 portrait pictures of individuals (6 male, 6 female) from the Basel Face Database (Walker et al., 2018). We randomly combined two male and two female faces into a permanent set of 4 individuals, which formed one group. This procedure allowed us to create three groups that consisted of four specific individuals each. The portraits were arranged randomly in a 2 × 2 grid (see Fig. 1 for an exemplary group picture and Appendix A for the overall setup).

The faces of depicted group members were manipulated in order to make the group objectively more diverse (compared to less diverse, labelled as homogeneous) on each dimension of the Big Five variables. To this end, we used the Basel Face Base Model (Paysan, Knothe, Amberg, Romdhani, & Vetter, 2009): Using the original pictures, we manipulated the pictures in regards to extraversion, conscientiousness, openness, agreeableness and neuroticism, applying changes of facial information so that persons depicted on the pictures would appear to look, for example, more or less extraverted. We then created groups by assembling four portrait pictures each. To create the more diverse groups, we systematically enhanced a specific personality dimension in the faces of two group members and reduced the same personality dimension in the other two. To create the less diverse group (labelled as homogeneous groups, which, however, always needs to be interpreted in relation to the more diverse group), all portraits of the group members were manipulated (enhanced or reduced) in the same direction in regards to the personality dimension (e.g., all were manipulated to appear more extraverted). Importantly, we calibrated the strength of the manipulation for the different personality traits to unity. Therefore, irrespective of which dimension was manipulated, the extent to which the faces change remains the same for all five personality dimensions.

Further information on which picture has been manipulated in which direction of the personality dimension in both the diverse or homogeneous condition can be found in Appendix A. For each of the three group pictures we thus created 10 different versions (i.e., a homogenous and a diverse version for each of the Big Five personality dimensions), while demographic attributes were kept constant, as the basic features of the group members did not change. In total, this resulted in 30 different group pictures that served as stimulus material for Study 1a. We predefined a specific arrangement of the 30 group pictures: We introduced a fixed alternation between the three groups (1,2,3), which was identical for all study participants, whereas the selection of the respective image for each group trial was random.

Further information on the size of the stimuli and background color of the survey for Study 1a and all further studies can be found in Appendix B.

4.1.4. Procedure

The participants were welcomed and learned that the aim of the study was to investigate how people form first impressions of groups based on portraits. After giving informed consent, participants learned that their task was to form an impression of groups and to evaluate how similar or different they perceived the group to be. Before starting with the main study, participants completed six exemplary trials to adjust to the task. These exemplary trials built on pictures with different faces taken from the Basel Face Database that were not used in the main study. Participants were asked to look at six different group pictures and rate whether they perceived the group to be similar or different on a 7-point Likert scale (1 = *similar*, 7 = *different*; representing the diversity rating). After the test phase, participants started working on the stimuli of the main study. Throughout 30 trials, the different group pictures were presented in a random order, alternating between the three groups (random picture of group one; random picture of group two; random picture of group three; for ten iterations in total). Participants were then asked to indicate their ID for compensation. At the end of the study participants indicated which devices they had used and whether they had encountered any problems with the appearance of the stimuli ("What kind of device did you use?": *Desktop PC, Notebook, Tablet, Smartphone, other device*; "Did you have to scroll down to see the whole page with the images and questions?": *never, sometimes, always*; "Was there any problem with an image not appearing?": *never, sometimes, always*). We asked for their gender, age, and language proficiency, as well as how carefully they answered our questions on a 7-point Likert scale (1 = *not carefully at all*, 7 = *very carefully*). Furthermore, participants were asked to indicate any reason not to use their data and to comment on the exact reasons, if so.[5] Finally, they were thanked for their participation and redirected to Prolific in order to get paid.

4.2. Results

After running the study, we noticed that a programming error caused the assembling of one of the group pictures to include three men and one woman (Group 2, neuroticism, diverse condition) instead of a 50:50 distribution as in all other pictures. We therefore omitted this trial from our data analysis.

Overall, participants descriptively seemed to be able to perceive differences in facial information of otherwise identical group pictures: Participants judged more diverse groups (compared to the less diverse groups) as more different. Table 1 summarizes the mean diversity evaluations of both the diverse and homogeneous version of the group pictures. In 9 out of 14 cases on group level, means differed in the predicted direction.

4.2.1. Analysis

On an average level, participants rated diverse group pictures as more diverse ($M = 3.96$, $SD = 1.13$) than homogeneous group pictures ($M = 3.77$, $SD = 1.12$), $t(99) = 4.60$, $p < .001$, $d = 0.46$. To disentangle whether this tendency was present for all five personality dimensions, we calculated separate *t*-tests for all of the five dimensions (see Table 2).

To investigate whether we could generalize our findings across participants as well as stimulus materials, we computed a linear mixed model (see Judd, Westfall, & Kenny, 2012) using the lme4 and lmerTest packages (Bates, Mächler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2017). In the analysis we included the fixed effects of the diversity manipulation (diverse or homogeneous group pictures, coded as 0.5 and − 0,5, respectively). We created the variable *trial information* (short: trial) as a product between group number and Big 5 trait, which thereby contained the ID of each group picture that was manipulated to be diverse versus homogeneous on the specific Big 5 dimension. Trial and participants were then included as random effects by modelling random intercepts for each variable and by-trials random slopes for the diversity manipulations. We omitted by-participants random slopes for the diversity manipulation as well as correlations between random effects, as including these led to convergence problems (Bates, Kliegl, Vasishth, & Baayen, 2015). Perceived diversity served as the dependent variable. The results of this analysis indicated that diverse groups were perceived as more diverse, even when including information of participants and trial into the model, $b = 0.16$, $SE\ b = 0.07$, $t(13.10) = 2.46$, $p = .029$. Additional analyses can be found in Appendix C.

[5] See Participants section. If people indicated reasons for data exclusion, we excluded them from the analyses.

Fig. 1. Two exemplary group pictures. For the left version extraversion is reduced in all faces (i.e., homogeneous group) whereas in the right version extraversion is enhanced for the two faces on the right and reduced for the two faces on the left (i.e., diverse group). Please note that in the study materials group members were randomly allocated to the positions within the 2 × 2 frame and the parallel alignment here was only chosen for the purpose of illustration. The original faces stem from the Basel Face Database (Walker et al., 2018).

Table 1
Mean (standard deviation) for all perceived diversity ratings across diversity manipulations, big five personality vectors, and groups.

Manipulated big five trait	Group	Diverse	Homogen.	
		M (SD)	*M (SD)*	
Conscientiousness	Group 1	4.01 (1.81)	3.71 (1.75)	*
	Group 2	3.08 (1.78)	3.30 (1.71)	
	Group 3	4.53 (1.62)	4.49 (1.58)	*
	Average	3.87 (1.31)	3.83 (1.27)	*
Agreeableness	Group 1	3.74 (1.76)	3.50 (1.60)	*
	Group 2	3.27 (1.59)	2.87 (1.52)	*
	Group 3	4.59 (1.68)	4.61 (1.64)	
	Average	3.87 (1.25)	3.66 (1.19)	*
Extraversion	Group 1	4.12 (1.70)	3.71 (1.71)	*
	Group 2	3.39 (1.65)	2.79 (1.59)	*
	Group 3	4.87(1.50)	4.70 (1.64)	*
	Average	4.13 (1.19)	3.73 (1.21)	*
Neuroticism	Group 1	3.70 (1.74)	3.81 (1.67)	
	Group 2	NA	3.31 (1.69)	
	Group 3	4.48 (1.60)	4.58 (1.68)	
	Average	4.09 (1.32)	3.90 (1.27)	*
Openness	Group 1	3.81 (1.73)	3.91 (1.75)	
	Group 2	3.25 (1.63)	2.95 (1.60)	*
	Group 3	4.63 (1.55)	4.33 (1.60)	*
	Average	3.90 (1.19)	3.73 (1.27)	*

Note. Mean differences that were descriptively in line with Hypothesis 1 (meaning that objectively diverse compared to homogeneous groups were perceived as more diverse) are highlighted with an asterisk. Inferential statistics can be found in Table 2.

4.3. Discussion

Study 1a was designed to test whether individuals' perceptions of diversity are informed by differences in facial information related to personality (Hypothesis 1*). The results from Study 1a provide support for this hypothesis. Overall, groups that were manipulated as more diverse (compared to less diverse, labelled as homogeneous) in regards to personality differences were perceived as more diverse. These results generalize across trials (computed as a product between group picture and Big 5 dimension) and participants.

When investigating effects for each personality trait individually (see also Appendix C), we did find stronger and weaker associations. Objectively, the manipulation strength for all the personality traits was the same. However, when asking participants whether they detected diversity in the groups, it was more likely that they did so when we manipulated diversity regarding extraversion, agreeableness, and eventually neuroticism (only when looking at *t*-test results), but not when we manipulated conscientiousness and openness.

On the one hand, these findings feed well into the discussion that objective and perceived diversity are not always associated and that associations may differ depending on the concepts involved. On the other hand, however, this finding regarding the differences between the Big 5 traits is of an exploratory nature and requires replication. Before, however, reporting the replication attempts of the effects of the manipulation of the Big 5 traits further, we describe Study 1b in which we tested whether the current findings can be extended to a different study setup.

5. Study 1b

Study 1a investigated diversity perceptions when participants were asked to rate one group picture after the other, without time constraints. Previous work investigating ensemble-coding (a visual mechanism to extract summary statistics from, for example, groups of faces; see Haberman et al., 2015; Phillips et al., 2018), however, has used a different approach. In these studies, two group pictures were presented sequentially and participants were asked for a comparison judgment. Participants first saw a target picture for a fixed amount of time (e.g., 2000 ms,

Table 2
Mean (standard deviation) for all perceived diversity ratings across diversity manipulations, big 5 personality vectors, and random vectors for studies 1a, 2a, and 2b.

	Study 1a			Study 2a			Study 2b		
Manipulated vector	Diverse	Homog.	Difference	Diverse	Homog.	Difference	Diverse	Homog.	Difference
	M (SD)	*M (SD)*		*M (SD)*	*M (SD)*		*M (SD)*	*M (SD)*	
Conscientiousness	3.87 (1.31)	3.83 (1.27)	$t(99) = 0.40$, $p = .687$, $d = 0.04$	3.73 (1.02)	3.64 (1.00)	$t(107) = 1.31$, $p = .194$, $d = 0.13$	3.73 (1.04)	3.83 (1.04)	$t(103) = -1.44$, $p = .152$, $d = -0.14$
Agreeableness	3.87 (1.25)	3.66 (1.19)	$t(99) = 2.41$, $p = .018$, $d = 0.24$	3.75 (0.94)	3.54 (0.95)	$t(107) = 2.48$, $p = .015$, $d = 0.24$	3.89 (1.12)	3.64 (1.03)	$t(103) = 2.79$, $p = .006$, $d = 0.27$
Extraversion	4.13 (1.19)	3.73 (1.21)	$t(99) = 4.74$, $p < .001$, $d = 0.47$	3.85 (1.04)	3.61 (1.02)	$t(107) = 2.86$, $p = .005$, $d = 0.28$	3.96 (1.06)	3.64 (1.07)	$t(103) = 3.84$, $p < .001$, $d = 0.38$
Neuroticism	4.09 (1.32)	3.90 (1.27)	$t(99) = 2.22$, $p = .029$, $d = 0.22$	3.66 (0.93)	3.71 (0.99)	$t(107) = -0.69$, $p = .490$, $d = -0.07$	3.71 (1.04)	3.84 (1.11)	$t(103) = -1.56$, $p = .123$, $d = -0.15$
Openness	3.90 (1.19)	3.73 (1.27)	$t(99) = 1.80$, $p = .074$, $d = 0.18$	3.74 (1.07)	3.55 (0.95)	$t(107) = 2.85$, $p = .005$, $d = 0.27$	3.78 (1.13)	3.65 (1.15)	$t(103) = 1.61$, $p = .111$, $d = 0.16$
Mean across Big 5	3.96 (1.13)	3.77 (1.12)	$t(99) = 4.60$, $p < .001$, $d = 0.46$	3.75 (0.85)	3.61 (0.85)	$t(107) = 3.50$, $p < .001$, $d = 0.34$	3.82 (0.94)	3.72 (0.95)	$t(103) = 2.46$, $p = .016$, $d = 0.24$
Random vector 1	–	–	–	3.67 (1.02)	3.60 (1.04)	$t(107) = 0.90$, $p = .372$, $d = 0.09$	3.69 (1.15)	3.68 (1.10)	$t(103) = 0.18$, $p = .855$, $d = 0.02$
Random vector 2	–	–	–	3.68 (1.01)	3.81 (0.96)	$t(107) = -1.56$, $p = .122$, $d = -0.15$	3.68 (1.10)	3.73 (1.12)	$t(103) = -0.62$, $p = .537$, $d = -0.06$
Random vector 3	–	–	–	3.85 (0.93)	3.57 (0.96)	$t(107) = 3.93$, $p < .001$, $d = 0.38$	3.98 (1.10)	3.69 (1.09)	$t(103) = 3.75$, $p < .001$, $d = 0.37$
Random vector 4	–	–	–	3.78 (0.97)	3.60 (0.93)	$t(107) = 2.20$, $p = .030$, $d = 0.21$	3.79 (1.18)	3.63 (1.09)	$t(103) = 2.18$, $p = .031$, $d = 0.21$
Random vector 5	–	–	–	3.75 (0.94)	3.54 (0.98)	$t(107) = 3.29$, $p = .001$, $d = 0.32$	3.71 (1.09)	3.72 (1.08)	$t(103) = 0.04$, $p = .965$, $d = 0.00$
Mean across random vectors	–	–	–	3.75 (0.83)	3.62 (0.84)	$t(107) = 3.30$, $p = .001$, $d = 0.32$	3.77 (1.00)	3.69 (0.96)	$t(103) = 2.28$, $p = .025$, $d = 0.22$

Note. None of the random vectors are identical in Studies 2a and 2b, so comparisons across studies are not possible for these variables.

see Phillips et al., 2018) and then a comparison picture. They were then asked to indicate whether the second picture was more or less diverse than the first. Participants were able to make correct and above chance level judgments for diversity regarding gender, race, or dominance (Phillips et al., 2018) and emotions (Haberman et al., 2015). Study 1b builds on this sequential setup and tests whether participants were able to make correct and above chance level choices when being asked to compare group pictures that differ in facial information associated with the Big 5 dimensions. Study 1b was preregistered: https://aspredicted.org/blind.php?x=/VD7_1DR.

5.1. Method

5.1.1. Participants

As approximation for a power analysis for mixed models, we computed a power analysis with G*Power (Faul et al., 2007) for a binomial test. This analysis indicates that a sample of 20 participants is required to test for differences from a constant (0.5) with an effect size estimate of $g = 0.30$, $\alpha = 0.05$, and a power of 0.80. To follow previous work (Phillips et al., 2018), we, however, aimed to collect 3000 data points (participants x trials). As the trial number is fixed to 30, we aimed to collect data from 100 participants to reach this objective. To buffer for exclusions, we added 10% to this number, aiming to collect data from a total of 110 participants.

We collected data from UK citizens for a study on group impressions via Prolific. One hundred and ten participants started the study and provided informed consent. One participant was excluded due to low carefulness ratings (< 3), and 5 participants indicated reasons not to use their data, resulting in an overall sample of 104 participants (17 male, 86 female, 1 non-binary; $M_{age} = 25.27$, $SD_{age} = 8.58$).

According to a sensitivity power analysis (calculations made using G*Power; Faul et al., 2007), a minimal effect size of $g = 0.14$ could be detected under standard criteria ($\alpha = 0.05$, $1 - \beta = 0.80$, two-tailed) with our sample of 104 participants. This, however, is only an approximation for the mixed model analyses reported below.

5.1.2. Design

Study 1b builds on a within design with 30 trials in which participants see the two versions of each group picture sequentially, one version is assembled as diverse the other as homogeneous (factor diversity). Using materials from Studies 2a and 2b, we used three different group pictures (factor group) that were manipulated on the five different Big 5 dimensions (factor Big 5 dimension). Choice accuracy serves as dependent variable and is coded as 1 for correct judgments of whether the comparison group is more or less diverse than the target group and 0 for incorrect judgment of whether the comparison group is more or less diverse than the target group.

5.1.3. Materials and procedure

Materials were taken from Studies 2a and 2b and we used the group pictures that were manipulated on the Big 5 dimensions. Having three different groups and 5 personality dimensions, this resulted in 15 different pictures overall with two versions (diverse vs. homogeneous), which could be displayed in two different orders (diverse vs. homogeneous version first). The stimuli, therefore, allowed us to create 30 different potential trials that were displayed in a randomized order (one was randomly selected as a practice trial, followed by 30 test trials). Due to stimuli-sampling, participants worked on one trial twice (the last of the 30 trials was one that they had been working on before, either in the practice or in the test trials).

Participants were welcomed to the study, provided informed consent, and were then asked to read the instructions. As in Philipps and colleagues (2018) we informed them that we were interested in judgments of diversity. We specified that we were interested in diversity in regards to personality. Participants learned that they would see two versions of a group and that the first one would be displayed for only 2 s. They would then need to judge whether the second version appeared more diverse in terms of personality than the first by pressing the keys “E” (less diverse) and “I” (more diverse) on their keyboards. Target pictures were presented for 2000 ms, followed by a blank screen for 100 ms, followed by the comparison picture which was displayed until participants made their decision. Participants then saw a blank screen for 1000 ms before continuing with the next trial.

Participants first worked on a practice trial containing one randomly chosen stimuli pair. After completing the practice trial, participants worked on 30 test trials. After completing all trials, we asked for demographics (gender and age), how carefully they had completed the study

(1 = *not carefully at all* to 5 = *very carefully*), whether there were reasons not to use their data, and whether they had any comments about the study.

5.2. Results

When looking at the 30 test trials, participants correctly identified whether the second version of the group picture was more or less diverse than the first version in on average 15.74 cases ($SD = 2.78$). The proportion of correct choices thereby lies at 52.47%.[6]

We computed a general linear mixed model to test whether choice accuracy differs from chance level. We did not include fixed effects, as we were especially interested in the model's intercept to determine whether choices differ from chance (see Phillips et al., 2018). We included random intercepts for participants and trials (combination of Big 5 dimensions, group picture, and presentation order; 30 levels). Results indicated that the intercept was significantly different from zero, $b = 0.10$, $SE\ b = 0.05$, $z = 2.16$, $p = .031$. Participants made 52.49% of correct choices based on model estimates, which differed from chance level. Changing the random effects from trials to trial ID (combination of Big 5 dimension and group picture as in previous and in the following studies; 15 levels) did not change the result pattern, $b = 0.10$, $SE\ b = 0.04$, $z = 2.22$, $p = .026$.

In an exploratory fashion, we further included the direction of change within the trial (coded as −0.5 when the diverse version was followed by a non-diverse version and 0.5 when the non-diverse version was followed by the diverse version). We further included by-participant random slopes for direction of change. Results corroborate the initial finding that the intercept differed from zero, $b = 0.10$, $SE\ b = 0.05$, $z = 2.18$, $p = .029$ and we did not find effects of direction of change, $b = -0.08$, $SE\ b = 0.09$, $z = -0.88$, $p = .377$. We further tested whether only looking at the first 29 evaluations would change the results (as trial 30 would use images that participants had already seen before, see Materials and Procedure), which was not the case: the intercept again differed from zero, $b = 0.09$, $SE\ b = 0.05$, $z = 2.01$, $p = .044$.

5.3. Discussion

Study 1b tested whether the association between facial variation on Big 5 variables and diversity perceptions could also be found in a different study setup typically used in the research on ensemble-coding (Haberman et al., 2015; Phillips et al., 2018). Looking at the results with a mixed model analysis indicated that participants were slightly but significantly better than chance level at identifying whether a second version of a group picture was more or less diverse than the previous version. The percentage of correct choices (52.49%) is in line with previous research investigating whether cues of social dominance impact perceived diversity in hierarchy (Phillips et al., 2018). Here, participants made 54% of correct choices. Higher levels of accuracy where achieved when identity relevant categories such as race (64% accuracy) or gender (60% accuracy) were varied (cf. Table 1; Phillips et al., 2018), which already indicates that differences in some attributes might be easier to detect and/or more likely to impact diversity perceptions than others (see Study 5 in this manuscript). All in all, the results of Study 1b speak to the generalizability of the findings of Study 1a to different research contexts.

While results from Study 1a and 1d provide first support for Hypothesis 1*, it remains unclear whether differences in facial information are considered more strongly for some of the Big 5 dimensions than for others or whether this finding resulted from chance. It is further unclear whether individuals considered differences in facial information in their diversity judgments *because* we had chosen personality-related variables (Big 5 dimensions) when manipulating the faces. Are these results driven by differences in perceptions of personality or more generally by differences in facial information alone that are not necessarily related to personality dimensions? Studies 2a and 2b were conducted to test the replicability of findings and to investigate whether facial information is taken into account generally or only when differences are related to Big 5 personality dimensions.

6. Study 2a

In preparation of Study 2a we created 15 novel vectors pointing in random directions in the face space. We aimed to keep intercorrelations between each random vector and the Big 5 vectors small. We selected the five vectors (out of the 15) that had the lowest maximum correlation across the Big 5 vectors ($rs < 0.07$). We then used the same group pictures as described above and applied the Big 5 (to replicate Study 1a) as well as the random vectors to create two additional versions of the same group picture: a more diverse and a less diverse (i.e., homogeneous) version. Study 2a was preregistered: http://aspredicted.org/blind.php?x=br96u2.[7]

6.1. Method

6.1.1. Participants

As this study is partly a replication of Study 1a, we aimed to obtain a similar sample size of $n = 100$. Due to the preregistered exclusion criteria, we added 10% to this number, resulting in a desired sample of 110 participants, to ensure that we arrive at the target number for our analysis. We collected data from UK citizens for a study on group impressions via Prolific. One hundred and eleven participants started the study and provided informed consent. Two participants were excluded as they indicated having technical problems with the images and one as they asked for exclusion explicitly, resulting in an overall sample of 108 participants (38 male, 70 female; $M_{age} = 34.59$, $SD_{age} = 13.45$).

According to a sensitivity power analysis (calculations made using G*Power; Faul et al., 2007), a minimal effect size of $d = 0.27$ could be detected under standard criteria ($\alpha = 0.05$, $1 - \beta = 0.80$, two-tailed) with our sample of 108 participants.

6.1.2. Design

Study 2a builds on a within design with four factors: diversity of group (manipulated as diverse vs. homogeneous), type of manipulation in faces (personality vs. random vectors), vector on which diversity was manipulated (Big 5 / random vectors 1–5), and group number (three different group constellations were used: Group 1, Group 2, Group 3). The resulting design is a 2 (diversity) x 2 (type of manipulation) x 5 (vector) within subjects design, with group number as a control factor. Perceived diversity served as a dependent variable.

6.1.3. Materials and procedure

Materials and procedure were identical to Study 1a, except a) the correction of the programming error in one stimulus and b) for the creation of a second set of picture versions by applying the five random vectors (an example of such an application can be found in Appendix D). Participants were therefore asked to rate the diversity of 60 images (instead of 30 in Study 1a). The specific manipulations for the diverse and homogeneous versions of the group pictures can be found in Appendix A.

6.2. Results

Overall, participants descriptively seemed to be able to perceive

[6] Computing a proportions test indicated that this number did not significantly differ from chance level, $\chi^2(1) = 0.01$, $p = .930$.

[7] We had an accidental deviation from the preregistration for Study 2a and 2b. As in Study 1a, we used a 7-point Likert-scale to assess the carefulness with which participants completed the study. The preregistration, however, states that the scale would range from 1 to 9. We use the same cut off value as preregistered (< 5) but wanted to make the difference in scale range transparent.

differences in facial information of otherwise identical group pictures. Table 2 summarizes the mean diversity evaluations of both the diverse and homogeneous version of the group pictures for each factor. On an average level, participants rated diverse group pictures as more diverse ($M = 3.75$, $SD = 0.82$) than less diverse (i.e., homogeneous) group pictures ($M = 3.62$, $SD = 0.82$), $t(107) = 4.58$, $p < .001$, $d = 0.44$. As a next step we tested whether this tendency was present for both faces manipulated on the Big 5 vectors and faces manipulated on the random vectors. Including diversity and type of manipulation as factors in a within-subjects ANOVA, we found a significant main effect of diversity, $F(1, 107) = 20.95$, $p < .001$, $\eta_G^2 = 0.01$, but no significant effect of type of manipulation, $F(1, 107) = 0.08$, $p = .776$, $\eta_G^2 = 0.00$, or the interaction between the two factors, $F(1, 107) = 0.09$, $p = .764$, $\eta_G^2 = 0.00$. Looking at the descriptive results, diverse (compared to homogeneous) group pictures were rated as more diverse when using Big 5 but also random vectors, see Table 2. To test whether differential effects found in Study 1a could be replicated, we calculated separate *t*-tests for all of the five Big 5 dimensions and random vectors (see Table 2).

We also used mixed models to analyze the data (as preregistered). We contrast coded diversity (−0.5 = *homogeneous*, 0.5 = *diverse*) and type of manipulation (−0.5 = *random vector*, 0.5 = *Big 5 vector*) and included both factors and their interaction as fixed effects into the model. As random effects we included a random intercept for participants and by-trial (the product between group number and Big 5 / random vector number to create unique trial IDs) random slopes for type of manipulation. We omitted other random effects and correlations between random effects due to convergence problems. Results corroborated the ANOVA analysis and indicated a significant main effect of diversity ($b = 0.13$, $SE\ b = 0.03$, $t(6341) = 4.30$, $p < .001$), but no significant main effect of type of manipulation ($b = -0.01$, $SE\ b = 0.27$, $t(28) = -0.03$, $p = .977$) or of the interaction between the two factors ($b = 0.02$, $SE\ b = 0.06$, $t(6341) = 0.25$, $p = .799$). Additional analyses can be found in Appendix C.

6.3. Discussion

The results of Study 2a replicate the main finding of Study 1a, namely that variation in faces impacts perceptions of diversity. Pictures that were created to be more or less diverse (while keeping demographic attributes such as gender, race, and age constant) were perceived as such by the participants, therefore providing support for Hypothesis 1.

Study 2a, furthermore, partly replicates some findings that are specific to the Big 5 traits. Looking at the impact of the diversity manipulation within the specific Big 5 factors, we again find a significant difference for extraversion and agreeableness, and a non-significant difference for conscientiousness. Different from Study 1a, the difference was significant for openness (when looking at *t*-test results), but not for neuroticism.

Last, Study 2a was also designed to determine whether the effect of manipulations in faces on diversity perceptions was related to the type of manipulation (as we used vectors that manipulated perceived personality regarding the Big 5 in faces). However, Study 2a shows that the effect is also present when using both the Big 5 vectors and random vectors that have been created in such a way that they are unrelated to the Big 5 personality vectors. Variance in facial information overall seems to impact diversity perceptions when other attributes are held constant (highlighting that Hypothesis 1* could be broadened to Hypothesis 1).

One caveat of the chosen procedure in Study 2a is that the vectors were created to be mathematically only weakly correlated with the Big 5 vectors. When applying these vectors to the faces, we did not test whether they might nonetheless impact Big 5 personality perceptions. Even though correlations between the random and Big 5 vectors were mathematically controlled, there could be some perceptual overlap. Study 2b was therefore conducted to control both mathematical and perceptual associations between Big 5 and random vectors. This setup provides an even stronger test of the importance of the type of manipulation in facial information on perceived diversity.

7. Study 2b

To obtain five random vectors that were also perceptually independent from the Big 5 vectors, we conducted an additional pretest. The 15 vectors created for Study 2a were applied to the twelve individual faces used in the Studies 1a, 1b, and 2a. For each face we created one version where the random vector was applied in the positive direction (+) and one with the vector applied in the negative direction (−). We then showed participants both versions of the faces and asked them which of the two versions looked more extraverted / open / agreeable / neurotic / conscientious. The Big 5 factors served as between subjects variables. Participants were only asked to judge one aspect out of five and we provided brief descriptions of the factors at the beginning of the study (adapted from Stolier, Hehman, Keller, Walker, & Freeman, 2018). One-hundred participants recruited via Prolific made these choices in 180 rounds (12 faces à 15 random vectors) distributed into two study blocks. We then looked at the results to see if the random vectors led to systematic differences in the perception of personality in faces. If the random vector was truly unrelated perceptually, choice proportions for each face should be 50:50 (chance level). We tested this across the twelve faces for each vector by looking at the absolute difference (weighted by the size of the sub sample for each of the Big 5 factors) in the distribution from chance level. The five random vectors that showed the lowest deviation scores were chosen for Study 2b (none of them were identical to the vectors used in Study 2a).

Study 2b used the newly chosen random vectors as a comparison to the Big 5 Factors. Participants in Study 2b were again presented with group pictures and asked for their diversity perceptions. The study was preregistered at https://aspredicted.org/blind.php?x=sk4x5e.

7.1. Method

7.1.1. Participants

As this study is a replication of Study 1a, we aimed to obtain a similar sample size of $n = 100$. Due to the preregistered exclusion criteria, we added 10% to this number, resulting in a desired sample of 110 participants. We collected data from UK citizens for a study on group impressions via Prolific. One hundred and ten participants started the study and provided informed consent. One participant was excluded due to low carefulness ratings (< 5), 2 participants were excluded as they indicated having technical problems with the images, and 3 as they asked for exclusion explicitly, resulting in an overall sample of 104 participants (32 male, 71 female, 1 no information; $M_{age} = 37.00$, $SD_{age} = 12.96$).

According to a sensitivity power analysis (calculations made using G*Power; Faul et al., 2007), a minimal effect size of $d = 0.28$ could be detected under standard criteria ($\alpha = 0.05$, $1 - \beta = 0.80$, two-tailed) with our sample of 104 participants.

7.1.2. Design

Study 2b builds on a within design with four factors: diversity of group (manipulated as diverse vs. homogeneous), type of manipulation (Big 5 personality dimensions vs. random vectors), vector on which diversity was manipulated (Big 5 / random vectors 1–5), and group number (three different group constellations were used: Group 1, Group 2, Group 3). The resulting design is a 2 (diversity) x 2 (type of manipulation) x 5 (vector) within subjects design, with group number as a control factor. Perceived diversity served as a dependent variable.

7.1.3. Materials and procedure

Materials and procedure were identical to Study 2a, except for the change in random vectors applied to the pictures (see Appendix A). Different from Study 1a and 2a, we slightly reworded the dependent

variable. In Study 2b, we asked participants to what extent they perceived the group members (previously: the group) to be similar or different as a measure for diversity perceptions.

7.2. Results

Overall, participants descriptively seemed to be able to correctly perceive differences in facial information of otherwise identical group pictures. Table 2 summarizes the mean diversity evaluations of both the diverse and homogeneous version of the group pictures for each factor. On an average level participants rated diverse group pictures as more diverse ($M = 3.79$, $SD = 0.95$) than homogeneous group pictures ($M = 3.70$, $SD = 0.94$), $t(103) = 2.85$, $p = .005$, $d = 0.28$. As a next step we tested whether this tendency was present for both types of manipulations used: the Big 5 vectors and the random vectors. Including diversity and type of manipulation as factors in a within-subjects ANOVA, we found a significant main effect of diversity, $F(1,103) = 8.13$, $p = .005$, $\eta_G^2 = 0.00$, but no significant effect of type of manipulation, $F(1, 103) = 1.69$, $p = .196$, $\eta_G^2 = 0.00$, or of the interaction between the two factors, $F(1, 103) = 0.05$, $p = .831$, $\eta_G^2 = 0.00$. Looking at the descriptive results, diverse (compared to homogeneous) group pictures were rated as more diverse when using Big 5 but also random vectors, see Table 2. To disentangle whether this tendency was present for all five personality dimensions, we calculated separate *t*-tests for all five dimensions (for results, see Table 2).

We further used mixed models to analyze the data. We contrast coded diversity (−0.5 = *homogeneous*, 0.5 = *diverse*) and type of manipulation (−0.5 = *random vector*, 0.5 = *Big 5 vector*) and included both factors and their interaction as fixed effects into the model. As random effects we included a random intercept for participants, but omitted a random intercept for trials (the product between group number and Big 5 / random vector number) in the model complexity reduction process (see Bates, Kliegl, et al., 2015). We further included by-trial random slopes for type of manipulation and the interaction between diversity and type of manipulation (correlations between random effects were omitted). Result, in tendency, corroborated the ANOVA analysis as we found an at least marginally significant main effect of diversity ($b = 0.09$, $SE\ b = 0.05$, $t(28.00) = 1.88$, $p = .071$), but no significant main effect of type of manipulation ($b = 0.04$, $SE\ b = 0.26$, $t(28.00) = 0.16$, $p = .875$) or of the interaction between the two factors ($b = 0.01$, $SE\ b = 0.10$, $t(28.00) = 0.09$, $p = .925$). When computing the same model as in Study 2a (random intercept for participants and by-trial random slope for type of manipulation), however, the main effect of diversity is significant ($b = 0.09$, $SE\ b = 0.03$, $t(6105) = 2.87$, $p = .004$). Additional analyses can be found in Appendix C.

7.3. Discussion

The results of Study 2b replicate the main finding of Study 1a and 2a, namely that ambiguous variations in faces impact perceptions of diversity, providing support for Hypothesis 1. Study 2b, furthermore, partly replicates the effects of the diversity manipulation within the specific Big 5 factors. Again, we found a significant difference for extraversion and agreeableness, and a non-significant difference for conscientiousness and neuroticism. Different from Study 2a (when looking at *t*-tests), the difference for openness was non-significant.

Last, Study 2b was designed to determine whether the effect of manipulations in faces on diversity perceptions was related to the type of manipulation (as we used vectors that manipulated perceived Big 5 dimensions in faces and vectors unrelated to the Big 5). However, 2b shows that the effect is also present when using random vectors that have been created in such a way that they are perceptually unrelated to the Big 5 personality vectors. Study 2b therefore provides additional support for the findings from Study 2a. It appears that effects of our manipulations on diversity perceptions are present across both types of manipulations (see main effect of diversity): Objective differences in facial information of group members can impact judgments regarding perceived diversity – and this effect occurs when manipulating facial information that is associated with Big 5 dimensions but also when manipulating facial information that is not related to the Big 5 dimensions. Study 2a and 2b further show that effect sizes are comparable for Big 5 manipulations (2a: $d = 0.34$, 2b: $d = 0.24$) and manipulations on random vectors (2a: $d = 0.32$, 2b: $d = 0.22$).

Based on these findings, we broadened Hypothesis 1* to Hypothesis 1 and refrained from the strong focus on differences in facial information relating to Big 5 personality dimensions (but see the General Discussion for a more thorough discussion). Hypothesis 1 then too relates to differences in ambiguous facial information more generally and not only facial information related to perceptions of Big 5 dimensions. However, we also call attention to the fact that when looking at the level of individual vectors, the pattern differs depending on the vector applied to the faces. Some of these patterns are also consistent across studies (e.g., extraversion and agreeableness) indicating that, even though changes are calibrated to be equally strong in all manipulations, some affect diversity perceptions more than others. While in the following Studies 3 and 4 we concentrate on the manipulation of extraversion, further discussions and research is required to better understand the differential patterns of differences in facial information on diversity perceptions.

8. Study 3

Studies 1a to 2b tested the impact of differences in facial information on diversity perceptions. In the next studies, we aimed to test how differences in facial information may not only impact perceptions but behavior, too. Looking at previous results, we concentrated on the factor with the strongest and most stable association between objective and perceived diversity: extraversion. We investigated whether individuals actively use this information when asked to assemble a more diverse versus a more homogeneous team, testing Hypothesis 2.

In Study 3, we asked participants to imagine being an HR-recruiter with the task of deciding between two possible new candidates for an existing team in their company. Participants were either instructed to choose the candidate that made the team more diverse or to choose the candidate that made the team more similar. To make their hiring choices, participants were asked to select one of two possible new candidates. Our hypothesis, the sample requirements, conditions, and planned analysis were preregistered: https://aspredicted.org/blind.php?x=87bx6e.

8.1. Method

8.1.1. Participants

Study 3 was conducted as an online experiment and advertised as a "group impressions" study on Prolific. The study took on average about 9 min to complete. The required sample size for Study 3 was computed using an a priori power analysis (Faul et al., 2007) assuming a medium effect size ($d = 0.5$) and a power of 0.80 ($\alpha = 0.05$), resulting in the requirement of 128 participants. We added 10% to that number while ensuring equal numbers in between subjects groups to ensure an adequate sample size even when applying our predefined exclusion criteria, resulting in a desired sample of 140 participants.

One hundred thirty-eight participants started the study and provided informed consent. Based on predefined exclusion criteria, we excluded 1 participant because they indicated reasons to not use their data, 2 participants because they rated themselves lower than 5 on our carefulness scale (1 = *not carefully at all*, 7 = *very carefully*) and another 5 participants because they reported display errors when looking at the presented group pictures. This procedure resulted in a final sample of 130 participants (44 male, 84 female, 2 not classified; $M_{age} = 36.47$, $SD_{age} = 12.42$). Within our sample, 66 participants were assigned to the homogeneous condition and the other 64 to the diverse condition.

According to a sensitivity power analysis (calculations made using G*Power; Faul et al., 2007), a minimal effect size of $w = 0.25$ could be

detected under standard criteria ($\alpha = 0.05$, $1 - \beta = 0.80$, two-tailed) with our sample of 130 participants.

8.1.2. Design

Study 3 builds on a design with one between subjects factor: Diversity instruction (assembling a diverse vs. homogeneous groups, between subjects). Participants' choice (candidate that resulted in a diverse vs. homogeneous team) served as a dependent variable, where choice for the diverse team was coded as 1 and choice for the homogeneous team was coded as 0.

8.1.3. Materials

We selected portraits of 4 males and 4 female face identities from the Basel Face Database (Walker et al., 2018) and reduced and enhanced extraversion, resulting in 16 portraits (i.e., 4 extraverted males, 4 extraverted females, 4 introverted males, 4 introverted females). We then created 24 male and 24 female trials as potential teams always consisted of either only males or females to circumvent the possibility of gender as a confound variable. In each trial, all four face identities from the same gender were used to create two group images. Two of the four face identities served as current team members, while the other two identities served as the applicants to this existing 2-person team. There are six different possibilities to select two of the four face identities per gender ($P(4,2) = 6$; i.e., IDs 1 & 2, IDs 1 & 3, IDs 1 & 4, IDs 2 & 3, IDs 2 & 4, IDs 3 & 4) as existing team members. The existing team members always had a similar "personality", that is, both appeared to be either introverted or extraverted, resulting in 12 possibilities to create existing teams. The personality of one applicant was similar to the existing team members (i.e., homogenous team), the personality of the other one was dissimilar (i.e., diverse team). Each face identity served as a current team member twice (once in the introverted version, once in the extraverted version). Adding the remaining two face identities as candidates results in 24 possible combinations for both genders (i.e., 48 combinations in total), since one of them can be portrayed as an introvert and the other as an extravert in one trial and vice versa in the other. Fig. 2 shows an example trial of one of the group situations.

Unfortunately, the same programming error as in Study 1a occurred,[8] resulting in an incorrect compilation of faces in two images (one female team image, one male team image), which were excluded from the data analysis. Overall, we analyzed 46 diversity choices of each participant.

8.1.4. Procedure

Participants were welcomed and learned that the aim of the study was to investigate how people form first impressions of groups based on portraits. After providing informed consent, the participants were instructed to imagine being an HR-recruiter in a company. Their task was to find the perfect new member for an existing team that consists of two members. Half of the sample was instructed to choose the candidate that they believed would make the group more similar on the basis of their first impression (homogeneity condition) and the other half was instructed to choose the candidate that they believed would make the group more different on the basis of their first impression (diversity condition). Before starting with the main study, participants completed two exemplary trials to adjust to the task. These exemplary trials were designed by compiling different pictures from the Basel Face Database (Walker et al., 2018) that were not used in the main part of the study. After the test phase, participants started working on the stimuli of the main study. Forty-eight trials were presented in a random order, while alternating between female and male trials. In each of these trials, participants were presented with two group pictures presenting the hypothetical group compilation consisting of the two current members and one of the two new candidates each. Then they were asked to choose between the candidates in alignment with their instructions.

Participants were asked for their Prolific ID, the device used, and whether they had any problems with the appearance of the stimuli ("What kind of device did you use?": *Desktop PC, Notebook, Tablet, Smartphone, other device*; "Did you have to scroll down to see the whole page with the images and questions?": *never, sometimes, always*; "Was there any problem with an image not appearing?": *never, sometimes, always*). We asked for their gender, age, and language proficiency as well as how carefully they answered our questions on a 7-point Likert scale (1 = *not carefully at all*, 7 = *very carefully*). Furthermore, participants were asked whether there was any reason not to use their data and if so, to comment on the exact reasons. Finally, they were thanked for their participation.

8.2. Results

Participants in the homogeneous condition chose the homogeneous option more often ($M = 27.94$, $SD = 5.88$) than the diverse option ($M = 18.06$, $SD = 5.88$), whereas participants in the diverse condition chose the diverse option more often ($M = 26.91$, $SD = 5.45$) than the homogeneous option ($M = 19.09$, $SD = 5.45$). When comparing the proportions in the two conditions, the tendency is visible, but does not reach significance, $\chi^2(1) = 2.68$, $p = .102$.[9]

To investigate whether our findings generalize across participants as well as stimulus materials, we computed a generalized linear mixed model. As a fixed effect we included the diversity manipulation, as random effects we included random intercepts for participants and trials as well as by-trial random slopes for the diversity manipulation. Choice of the diverse (coded as 1) versus homogeneous team (coded as 0) served as the dependent variable and we used a generalized mixed model (logit) to analyze the data. Results showed that the diversity instruction (compared to the homogeneous instruction) led to a higher probability of participants selecting diverse groups, $b = 0.94$, $SE\ b = 0.26$, $z = 3.63$, $p < .001$.

8.3. Discussion

After having shown that individuals' perception of diversity captures differences in facial information, Study 3 focused on individuals' behavior, meaning their ability to create more diverse or homogeneous teams in regards to differences in faces. The results of Study 3 provide support that individuals can, if instructed to do so, use their perceptions of differences regarding facial information to actively assemble more diverse or more homogeneous teams, thereby providing support for Hypothesis 2.

Building on these findings, we next aimed to investigate whether individuals' diversity beliefs (the extent to which they see value in diversity versus homogeneity of a group; van Knippenberg & Haslam, 2003) significantly predict their choice behavior when assembling teams differing in regards to their members' personality. We hypothesized that diversity beliefs predict applicant selection such that individuals who believe in the value of diversity are more likely to assemble diverse rather than similar teams in regards to facial information than individuals who believe in the value of homogeneity (see Hypothesis 3). Study 4 was designed to test this assumption.

9. Study 4

Diversity beliefs refer to beliefs about the value of diversity for work

[8] As Studies 3 and 4 followed after Study 1a, they are also affected by the same programming error. Studies 1b, 2a, 2b, and 5 followed later, and the error was corrected by then.

[9] In the preregistration, we originally registered to compare choice proportions with a *t*-test. We, however, later noticed that this is not the best approach to analyze the count data and therefore report the proportions-test above. When comparing average choices with a *t*-test, results indicate that conditions differed significantly; $t(127.8) = -8.90$, $p < .001$, $d = -1.56$.

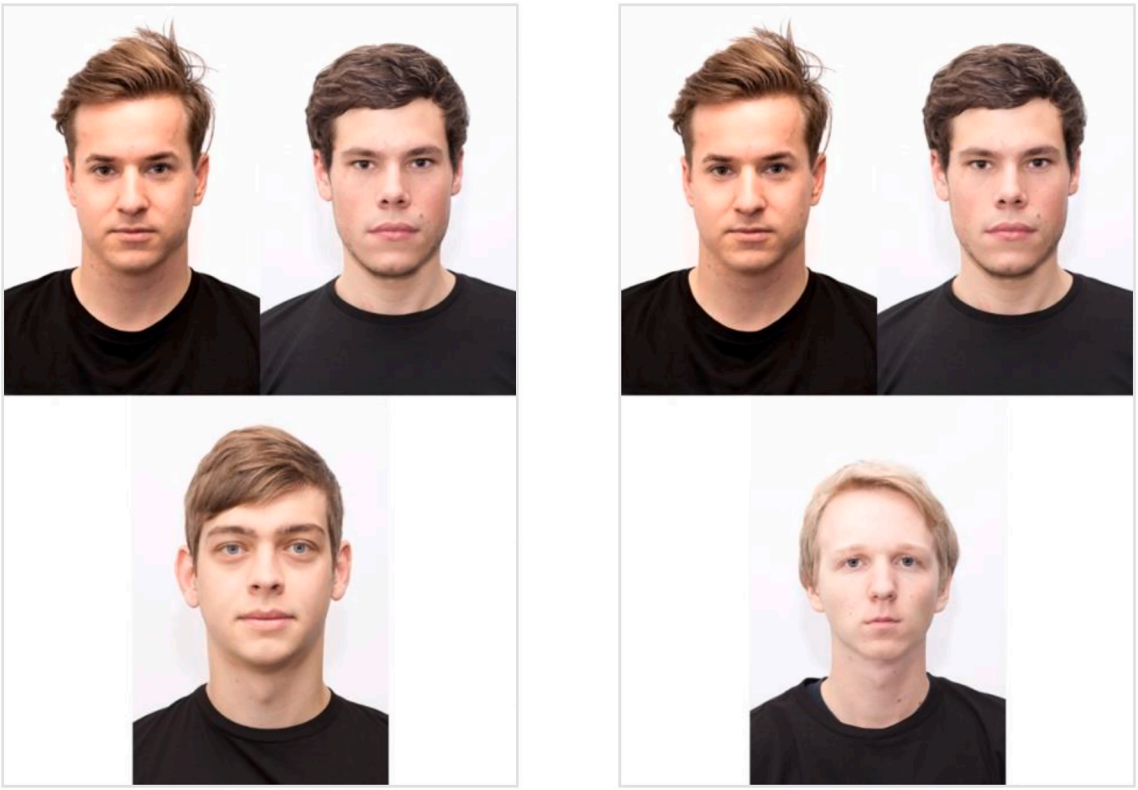


Fig. 2. This image illustrates one trial (group situation). The upper four portraits represent the two current team members, while the lower two portraits represent the two applicants. Participants saw the two potential teams that they could assemble by choosing the left or the right applicant for the team. In this case both current team members look extraverted. The left applicant is also manipulated to look extraverted, while the right candidate is manipulated to look introverted. Choosing the left candidate results in a homogenous team, choosing the right candidate results in a diverse team.

group functioning (van Knippenberg & Haslam, 2003). Diversity beliefs describe whether or not individuals actually prefer to work with different (versus similar) others. The extent to which individuals see value or threat in diversity (compared to value or threat in homogeneity) might critically impact the consequences of diversity. Previous research has shown that work group diversity and group identification are more positively related the more individuals believe in the value of diversity (van Knippenberg, Haslam, & Platow, 2007). Diversity beliefs further moderate the negative association between perceived diversity and team identification as well as the positive association between perceived diversity and relationship conflict (Hentschel, Shemla, Wegge, & Kearney, 2013).

Diversity beliefs therefore may predict whether individuals would choose a more diverse or more homogeneous team. However, it is an open question whether diversity beliefs are predictive for the selection of groups that differ in facial features. In Study 4, we focused on individuals' diversity beliefs and their impact on choice behavior when confronted with facially diverse or homogeneous teams. We hereby test Hypothesis 3 and aim to understand whether individuals who see more value in diversity are also more likely to select facially diverse compared to homogeneous teams.

In Study 4, we again asked participants to imagine being an HR-recruiter with the task of deciding between two possible new candidates for an existing team in their company. However, we aimed to create a more ambiguous situation, in which selection strategies were less clear and individuals needed to find an individual strategy to solve the selection task. To make their hiring choices, we used the same procedure as in Study 3. Thus, participants were asked to select one of two possible new candidates that would join an existing team. Our hypothesis, the sample requirements, conditions, and planned analysis were preregistered: https://aspredicted.org/blind.php?x=s8bz8d.

9.1. Method

9.1.1. Participants

Study 4 was conducted as an online experiment and advertised as a "group impressions" study on Prolific. The study took on average about 11 min to complete. The required sample size for Study 4 was computed using an a priori power analysis (Faul et al., 2007) assuming a medium effect size ($d = 0.5$) and a power of 0.80 ($\alpha = 0.05$, one-sided testing), resulting in the requirement of 64 participants. We added 10% to that number to ensure an adequate sample size even when applying our predefined exclusion criteria, resulting in a desired sample of 71 participants.

Seventy-one participants started the study and provided consent. Based on predefined exclusion criteria, we excluded 1 participant due to low carefulness ratings (<5) and 1 participant because they reported display errors when looking at the presented group pictures. This procedure resulted in a final sample of 69 participants (28 male, 40 female, 1 not classified; $M_{age} = 33.25$, $SD_{age} = 11.04$).

According to a sensitivity power analysis (calculations made using G*Power; Faul et al., 2007), a minimal effect size of $r = 0.29$ could be detected under standard criteria ($\alpha = 0.05$, $1 - \beta = 0.80$, one-tailed) with our sample of 69 participants.

9.1.2. Design

Study 4 builds on a correlational design: Diversity beliefs (as a continuous attitude measure of whether individuals see value in diversity versus homogeneity, between subjects) served as a predictor variable. Participants' choice (diverse versus homogeneous candidate) served as a dependent variable, where choice for the diverse team was coded as 1 and choice for the homogeneous team was coded as 0.

9.1.3. Materials

The materials in Study 4 were identical to the materials used in Study 3. To assess individuals' diversity beliefs, we used four statements related to diversity (Homan, Greer, Jehn, & Koning, 2010; Homan, van Knippenberg, Van Kleef, & De Dreu, 2007) to which individuals were asked to rate their agreement on a seven-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*). An exemplary item was "I enjoy working in diverse groups". As in Study 1a and 3, the same programming resulted in a wrong compilation of faces in two images (one female team image, one male team image), which were excluded from the data analysis. Overall, we analyzed 46 diversity choices of each participant.

9.1.4. Procedure

The procedure of Study 4 closely resembles the procedure of Study 3 with the following differences. Participants were instructed to imagine being an HR-recruiter in a company. Their task would be to find a new member for an existing team that consists of two members. However, to create a more ambiguous situation in which the selection strategy could mirror individuals' diversity preferences, we told individuals that the company they were working for had three core values, which they should consider when making their hiring choices. These core values were "Respect for colleagues, clients, and all people met along the course of the work", "Diversity as a crucial aspect to nurture creativity", and "Partnership as a foundation for the work with clients".

After working on the 48 selection trials, participants were asked to indicate to what extent the respective core values had influenced their decisions and rated the impact of each of the three core values on a seven-point Likert scale (1 = *no influence*, 7 = *strong influence*). Next, participants worked on the items on diversity beliefs, before completing questions on technical challenges as well as demographics (see Procedure of Study 3 for further details).

9.2. Results

On average, participants assembled a diverse team in 22.39 trials ($SD = 4.45$) and a homogeneous team in 23.61 trials ($SD = 4.45$). We calculated a Pearson correlation test to investigate the association between diversity beliefs and the number of trials in which a diverse team was selected. Results indicate a positive correlation between diversity beliefs and number of trials in which participants chose a candidate that would result in a more diverse team; $r(67) = 0.34$, $p = .005$.

To investigate whether our findings generalize across participants as well as stimulus materials, we computed a generalized linear mixed model for binary outcomes. We included diversity belief as a fixed effect. As random effects we included random intercepts for participant and trials. Choice of the diverse (coded as 1) versus homogeneous team (coded as 0) served as the dependent variable. Our results indicate that diversity beliefs predicted the likelihood that a diverse team was assembled, $b = 0.10$, $SE\ b = 0.03$, $z = 2.95$, $p = .003$.

Exploratorily, we also investigated whether individuals could correctly indicate their reasons for choosing the teams. To this end, we investigated whether individuals who stated that the core value of diversity had a stronger influence made more choices in favor of diversity. The correlation between self-reported strength of influence and choice behavior was positive, $r(67) = 0.26$, $p = .031$.

9.3. Discussion

Study 4 was designed to investigate the impact of individuals' diversity beliefs on their choice behavior when assembling more diverse or homogenous teams in regards to facial variations that mirror differences in extraversion. The results support Hypothesis 3 in showing that individuals who see value in diversity are more likely to assemble more diverse teams than individuals who see value in homogeneity. Although differences between group members are ambiguous, individuals' beliefs matter and predict their selection strategies and preferences in the task setting.

10. Study 5

In Studies 3 and 4, we tested whether differences in faces would impact diversity-related decision making when other variables such as gender and race are kept constant. It is, however, interesting to understand whether differences in faces may or may not play a role, when other variables that are more closely linked to the perceptions of diversity (Harrison et al., 2002) differ, too. To investigate this research question, we created a study similar to Studies 3 and 4 and asked participants to imagine being an HR-recruiter with the task of deciding between two possible new candidates for an existing two-person team in their company. Participants learned at the beginning of the study that one of the core values of the company was diversity. They read that the company truly believed that having a group of people with different perspectives, ideas, approaches, personalities, and working styles enhanced creativity, innovativeness, and performance of the group and thereby of the entire organization. Their choices, therefore, should be led by these considerations.

The setup of the group in Study 5 differed from the previous studies: The two-person team always consisted of one female and one male person and both were manipulated to appear extraverted or introverted. The team was therefore diverse in regards to gender and homogeneous in regards to perceived extraversion. The two candidates in Study 5 were also conceptualized differently: One candidate was female and one was male, and one was manipulated to appear extraverted and the other one introverted. Choosing one candidate would result in a more homogeneous team regarding extraversion, choosing the other in a diverse team.

This setup resulted in two types of trials: One, in which the choice of the male candidate would result in higher diversity, as the male candidate's face was manipulated in the same direction as the team in regards to extraversion and the female candidate's face in the opposite direction (Type 1). In the other trials the opposite was true and the choice of the female candidate would result in higher diversity regarding facial features associated with extraversion (Type 2). Note that the resulting diversity regarding gender of the three-person team would objectively be the same regardless of which candidate would be chosen (as it would always result in a 2:1 distribution). However, subjectively a two women-one man versus two men-one women distribution could result in different levels of *perceived* diversity, as choosing a female candidate could increase perceived gender diversity more than choosing a male candidate, as women are eventually more likely to be underrepresented in parts of the workforce (see Phillips et al., 2018). This means that in Type 1 trials personality and demographic-associated diversity is misaligned (meaning that the choice of the male candidate would lead to the facially diverse team) and in Type 2 trials personality and demographic-associated diversity is aligned (meaning that the choice of the female candidate would lead to the facially diverse team).

This study setup allowed us to test whether participants would choose the facially different candidate when teams and candidates also differed on gender. It also allowed us to test whether this tendency differed for the type of trial that we created. Our hypothesis, the sample requirements, conditions, and planned analysis were preregistered: https://aspredicted.org/blind.php?x=/NMP_AXZ.

10.1. Method

10.1.1. Participants

Study 5 was conducted as an online experiment and advertised as a "Group impressions" study on Prolific. The study took on average about 11–12 min to complete. As approximation for a power analysis, we used Pangea (Westfall, 2016) to estimate that a participant sample size of $n = 90$ completing 144 trials is required to detect small to medium fixed effects ($d = 0.30$) with a power of 0.80 (α -level = 0.05). Due to the predefined exclusion criteria, we added ~10% to this number (resulting in a number of 100 participants), to ensure we arrive at a minimum of 90 participants for our analysis. Pangea (Westfall, 2016), however, is

created for power analyses for models with continuous dependent variables, whereas Study 5 has a categorical dependent variable (choice of diverse versus homogeneous team). To corroborate the analysis, we further computed a power analysis with G*Power (Faul et al., 2007) for a binomial test. This analysis further indicated that a sample of $n = 20$ participants is required to test for differences from a constant (0.5) with an effect size estimate of $g = 0.30$, $\alpha = 0.05$, and a power of 0.80. A sample size of 90 would allow to detect effects of $g \geq 0.15$ with a power of 0.80.

One hundred participants started the study and provided consent. Based on predefined exclusion criteria, we excluded 2 participants due to low carefulness ratings (<5) and 3 participants because they reported display errors when looking at the presented group pictures. This procedure resulted in a final sample of 95 participants (32 male, 62 female, 1 non-binary / third gender; $M_{age} = 32.12$, $SD_{age} = 10.71$).

According to a sensitivity power analysis (calculations made using G*Power; Faul et al., 2007), a minimal effect size of $g = 0.15$ could be detected under standard criteria ($\alpha = 0.05$, $1 - \beta = 0.80$, two-tailed) with our sample of 95 participants.

10.1.2. Design

Study 5 builds on a within design with one within subjects factor: type of trial (Type 1-misaligned where the facially different candidate was male vs. Type 2 where the facially different candidate was female). We also refer to the trials as Type 1-misaligned (as personality and demographic associated diversity are misaligned) and Type 2-aligned (as personality and demographic associated diversity are aligned). Participants' choice (candidate that led to a diverse vs. homogeneous team in regards to extraversion) served as a dependent variable, where choice for the diverse team was coded as 1 and choice for the homogeneous team was coded as 0. Choice behavior was investigated across all 144 trials, but we also preregistered to look at the first 48 choices only, to make results comparable to Studies 3 and 4 and eventually mitigate the impact of study fatigue on our results. In an exploratory fashion, we also recoded choice behavior to test whether participants were more likely to choose the female (coded as 1) compared to the male (coded as 0) candidate.

10.1.3. Materials

The materials in Study 5 were similar to the materials used in Studies 3 and 4 (see Fig. 2). We used three female and male faces from the Basel Face Database (Walker et al., 2018) that were manipulated to appear more extraverted and more introverted (two versions for each face). We then created teams that fulfilled the criteria described in the study introduction: We created two-person teams consisting of a female (displayed left) and male (displayed right) candidate and added one male and one female candidate of which one was manipulated to appear extraverted and one as introverted. Creating all possible combinations resulted in a total of 144 decision pairs and therefore 144 trials for the study.

10.1.4. Procedure

The procedure of Study 5 closely resembles the procedure of Studies 3 and 4 with the following differences. Participants were instructed to imagine being an HR-recruiter in a company. Their task was to find a new member for an existing team that consists of two people. We told participants that one of the core values of the company they were working for was diversity but specified that diversity referred to different perspectives, ideas, approaches, personalities, and working styles (see above). Participants were asked to keep this in mind when making their hiring choices.

Participants first completed two practice trials (taken from Studies 3 and 4, meaning that here, there were no gender differences within the teams), before proceeding to the main study. After working on the 144 selection trials, participants were asked to indicate how strongly the company's core value had influenced their decisions (7-point Likert scale; 1 = *no influence*, 7 = *strong influence*). Next, participants completed questions on technical challenges as well as demographics (see Procedure of Study 3 for further details).

10.2. Results

10.2.1. Preregistered analyses

On average, participants assembled a diverse team in 72.40 ($SD = 14.09$) out of 144 trials. In Type 1-misaligned trials (personality and demographic associated diversity was misaligned), participants assembled a diverse team in 29.47 ($SD = 13.49$) out of 72 trials, in Type 2-aligned trials (personality and demographic associated diversity was aligned), they assembled a diverse team in 42.93 ($SD = 13.76$) out of 72 trials. We calculated a proportion test to investigate whether choice proportions differed from chance level, which was not the case (overall $\chi^2(1) = 0.00$, $p = 1.00$; Type 1 trials $\chi^2(1) = 2.02$, $p = .156$; Type 2 trials $\chi^2(1) = 2.29$, $p = .130$.

To test our assumptions across participants and trials, we computed a generalized linear mixed model for binary outcomes. We included type of trial as a fixed effect (contrast coded with −0.5 = *Type 1-misaligned trials* and 0.5 = *Type 2-aligned trials*). As random effects we included random intercepts for participant and trials and by-participant random slopes for type of trial. Choice of the facially diverse (coded as 1) versus homogeneous team (coded as 0) served as the dependent variable. Next to the fixed effect, we were especially interested in the intercept, as it can be used to determine whether choices differ from chance (see Phillips et al., 2018). Our results show a non-significant intercept, $b = 0.01$, $SE\ b = 0.06$, $z = 0.15$, $p = .878$, and a significant effect of type of trial, $b = 1.11$, $SE\ b = 0.24$, $z = 4.60$, $p < .001$, indicating that participants were more likely to assemble the facially diverse team in Type 2-aligned compared to Type 1-misaligned trials. This finding was corroborated when only looking at participants' first 48 choices (intercept: $b = 0.04$, $SE\ b = 0.05$, $z = 0.79$, $p = .430$; type of trial: $b = 1.02$, $SE\ b = 0.21$, $z = 4.79$, $p < .001$).

10.2.2. Exploratory analyses

In an exploratory fashion, we extended the mixed model reported above and added participants' self-indicated influence of the core value diversity on their choices (*z*-standardized). Diversity influence significantly predicted diversity choices, $b = 0.15$, $SE\ b = 0.05$, $z = 2.90$, $p = .004$. In a next step, we included an interaction term for type of trial x diversity influence and find that the association between diversity influence differs for type of trial, $b = 0.58$, $SE\ b = 0.23$, $z = 2.52$, $p = .012$. To disentangle this interaction effect, we computed a mixed model for the different type of trials with influence of diversity as fixed effects and random intercepts for participants and trials. Results indicate a negative but non-significant association between diversity influence and choices for Type 1-misaligned trials ($b = -0.14$, $SE\ b = 0.12$, $z = -1.14$, $p = .253$) and a positive association for Type 2-aligned trials ($b = 0.40$, $SE\ b = 0.11$, $z = 3.60$, $p < .001$). For further analyses, see Appendix C.

10.3. Discussion

Study 5 was designed to investigate the impact of facial variation when other attributes such as gender differ, too. Participants were asked to select candidates that were similar or different in facial information (related to extraversion) compared to the two-person team and were female versus male. Results indicate that participants' choice proportion for the candidate that would result in a facially diverse team did not differ from chance level. Instead, participants were more likely to choose the female compared to male candidate (see the difference between Type 2-aligned and Type 1-misaligned trials), alluding to the idea that including women in teams might increase perceived diversity more due to their underrepresentation in certain areas of the workforce (see Phillips et al., 2018). Results indicate that when participants are asked to assemble a diverse team, facial information may be taken into account

(see Studies 3 and 4), but this effect may deteriorate if other, more salient attributes such as gender (which is also more strongly associated with diversity definitions) differ, too.

Study 5, however, may have been a conservative test to investigate whether facial information is taken into account when other, more salient attributes differ, too. When informing participants about the company's objective, we described that the company truly believed that having a group of people with different perspectives, ideas, approaches, personalities, and working styles enhanced creativity, innovativeness, and performance. We did mention differences in personality in this description, however, we offered other rationales, too, which may be conceptually associated with gender diversity (e.g., different perspectives). In a follow -up study (Study 5b, reported in Appendix E), we therefore added a systematic variation to the instructions to the study: Half of the participants were asked explicitly to focus on gender-related diversity and the other half on personality-related diversity. When looking at choice proportions, participants were more likely to choose the facially diverse team when being instructed to focus on personality than on gender. Participants' focus is important and while it may be the case that in everyday life people are more likely to focus on gender diversity, facial information can be taken into account if the context requires it.

Next to highlighting the importance of participants' focus, another reason could explain why in Study 5 participants were more likely to attend to gender: One could speculate that differences in gender and in manipulated extraversion might not be comparable in strength, and gender differences could therefore be more salient in the pictures. We discuss opportunities for future research in the next section.

11. General discussion

This project investigates perceptions of diversity in regards to ambiguous cues such as differences in facial information. More specifically we tested whether a) objective diversity in facial information (reflecting perceived personality but also variation that is not related to the Big 5 dimensions) impacts perceived diversity and b) whether there are boundary conditions under which people are more versus less likely to take diversity in facial information into account.

The results of this manuscript support the hypothesis that differences in facial information, under certain conditions, may impact diversity perceptions and choices. In Study 1a we were able to show that two groups consisting of the same individuals were perceived as more or less diverse, depending on whether the faces signaled different or similar levels of specific personality dimensions. In other words, differences in facial information related to perceptions of personality led to significant differences in perceived diversity. Study 1b showed that effects of the Big 5 differences on perceived diversity can also be found when asking participants to make comparison choices under time constraints. Participants were slightly better than chance level at identifying whether the latter version of a group picture was more or less diverse than the former version of the group picture. Interestingly, participants showed this above chance level performance in a context where they had 2 s to look at the first picture and then compare their impression to another picture without knowing on what specific attributes they could focus.

Study 2a and 2b replicated and extended these findings by showing that the pattern was not specific to Big 5 personality dimensions but also occurred when using vectors unrelated to the Big 5 to create more diverse or homogeneous groups in regards to their faces.

Study 3 then provided support that individuals use their diversity perceptions (regarding extraversion) to assemble more diverse or homogeneous teams. Importantly, participants were not instructed to use perceptions of extraversion or more generally personality, yet succeeded in creating more or less facially diverse groups in accordance with task instructions.

The results of Study 4 then showed that individuals' diversity beliefs predicted choice behavior, highlighting that when they saw value in diversity (compared to homogeneity) they were more likely to assemble more facially diverse teams (the differences in facial information again representing perceived extraversion). Thus, without clearly instructing participants on what kind of strategy they should use to assemble teams, differences in facial information were spontaneously considered the more participants believed in the value of diversity.

Study 5 then tested whether facial diversity impacted choice behavior when other and eventually more salient attributes such as gender differ too. Choice proportions for the candidate that increased facial diversity of the team did not differ from chance level; instead participants displayed a preference for female over male candidates. Including gender differences, therefore, deteriorated the effect of facial differences on choice behavior. When looking at choice proportions of female candidates in an exploratory fashion (see Appendix C), however, results indicated that facial diversity may still have an impact: When looking at the association between self-indicated diversity influence and choice behavior, this link was stronger in trials where the female (vs. male) candidate would also increase the facial diversity of the group.

11.1. Theoretical contributions of the present findings

These results contribute to a better understanding of perceived diversity (Shemla & Meyer, 2012; Shemla, Meyer, Greer, & Jehn, 2016) and provide an answer to the question of whether differences in facial information may influence diversity perceptions. Results show that when keeping information on demographic attributes such as race and gender (which are often associated with diversity) constant by using the exact same persons in pictures of teams, subtle differences in facial information (that can but do not need to be related to the Big 5, see Studies 2a and 2b) are detected and used to form diversity perceptions.

Looking more closely at the specific dimensions manipulated in the studies, effects differed among the random vectors, showing that effects of some of the manipulations on diversity perceptions were similar in size to extraversion and agreeableness effects (for example, vector 3 (V12) in Study 2a), whereas others resulted in opposing or null effects. Although we ensured that the random vectors were largely independent of the Big 5 vectors, it may still be the case that some of them resulted in facial changes that were meaningful to the participants and that participants considered when forming their perception judgments. One could speculate that one vector may have changed other personality-relevant information, holistic facial features, such as attractiveness, babyfaceness, masculinity/femininity, or specific facial features such as size of eyes or contrast of eyes. To illustrate potential effects, we provide an example of random vector 3 (V12) in Appendix D.

When further looking at the levels of the Big 5, differences are apparent as well. Differences in extraversion and agreeableness predicted perceived diversity, whereas the effects of the other dimensions on diversity perceptions were less clear. One could speculate that extraversion and agreeableness perceptions could be related to perceived affect or emotions (even if facial changes simply resemble emotional expressions; see Jaeger & Jones, 2021). Perceived affect, further, seems to be one of the facial dimensions that people rely on most when judging others (Jaeger & Jones, 2021; see also the literature on the emotion overgeneralization effect; Zebrowitz & Montepare, 2008) and other work has shown that differences in emotions can impact diversity judgments (Haberman et al., 2015). As no specific instructions have been given in most of the studies, participants may have spontaneously attended to these features, thereby finding it easier to detect differences in extraversion and agreeableness.

This research shows that facial information may, under certain controlled circumstances, impact diversity perceptions - but this association may not hold under other circumstances. In Study 5 differences

in facial information were not the only cue available and were complemented by differences in gender. Here, results do not show support for the hypothesis that differences in facial cues impact diversity choices – instead, participants were more likely to choose female instead of male candidates, highlighting an interesting boundary condition of the effect. In everyday life, there are situations in which identity related attributes such as gender are kept constant (e.g., sports teams, pictures on dating platforms when people are interest in one gender only, oftentimes management teams), yet more often attributes such as gender and race will differ too. Here, the impact of differences in facial information may be much smaller or negligible (unless there might be a reason to pay particularly strong attention to it, due to task requirements or instructions, see follow-up Study 5b in Appendix E).

Last but not least, this research focuses on the importance of individuals' diversity beliefs in regards to choice behavior, showing that individuals assemble a more diverse team when they themselves see value in diversity. Perceptions might be the basis for choice behavior, but the more important determinant for building diverse teams could be how much value individuals place on diversity in general and on certain attributes in particular.

11.2. Limitations and future research

The present set of studies were all conducted in an experimental setting, in which participants were asked to evaluate or choose different candidates based on visual stimuli (portraits). Across studies, we did not instruct participants to attend to personality or to specific Big 5 dimensions, therefore they might have used information that they spontaneously rely on when making judgments (e.g., perceived emotions; see Jaeger & Jones, 2021; or trustworthiness, see Klapper et al., 2016). Eventually adding specific instructions on what people should attend to or operationalizing the dependent variable as personality diversity or more specifically diversity in extraversion could result in stronger effects and could increase differences for group pictures manipulated on the Big 5 compared to the group pictures that were manipulated by using vectors that were not related to the Big 5. Future work could test how specific instructions and variable operationalizations could guide participants' focus and impact diversity perceptions. The follow-up Study 5b (Appendix E) shows that instructing participants to focus on gender- versus personality-diversity impacts the choices of teams that participants make.

Further work could also look at the context of decision making (see Todorov, Said, Engell, & Oosterhof, 2008) and introduce contexts where certain variables may be more or less important (e.g., work vs. leisure groups). Here, research could investigate whether participants are able to attend more or less to facial variation, depending on the association between dimensions and importance in the specific context.

On the theoretical side, future work could also test whether the results of Studies 3 and 4 are replicable with other manipulations such as variation in perceived agreeableness or also in random vectors applied to the faces. Other work has identified trustworthiness and dominance (Oosterhof & Todorov, 2008; Todorov et al., 2008), but also attractiveness (see Sutherland et al., 2013) as dimensions that individuals use to evaluate others. These dimensions could be used to test whether diversity in facial information would result in perceived diversity as well. The work by Phillips et al. (2018) may already speak to this question, as the authors show that differences in dominance in faces impact perceptions of variance in hierarchy.

Future research could also investigate more specific diversity beliefs regarding perceived personality traits, such as extraversion in Studies 3 and 4 or agreeableness in the suggested research. It could be interesting to ask participants not only whether they see value in diversity in general, but what they believe is the value in diversity in extraversion or agreeableness. Eventually, beliefs could differ, as one may assume that having both introverts and extraverts in a team could sound promising, whereas it may not sound as promising to include a disagreeable person in a team (Rudert, Keller, Hales, Walker, & Greifeneder, 2020). Disentangling individuals' beliefs regarding specific personality dimensions might further predict choice behavior when assembling teams.

Apart from Study 5, we kept team variables constant except for facial information with which we manipulated objective diversity versus homogeneity of a team. This setup allowed us to carefully test the impact of variation of facial information in a very controlled setting. The downside of this approach is, however, that it is unclear whether the cues influence diversity perceptions and choices when groups differ on other and more salient variables, too. Study 5 suggests that the impact of facial variation deteriorates when differences in gender are present. In a next step, the designs of Studies 3, 4, and 5 could be combined into one study, testing in the same setting whether differences in facial information influence diversity choices only in conditions where gender does not differ (mimicking Studies 3 and 4) but do not do so when gender differs (mimicking Study 5).

Further, future research could test the impact of facial variation when meaningful and salient attributes such as gender (see Study 5) or race differ, but also when less salient attributes such as the color of clothing varies. Alternatively, facial information could be made more salient by using stronger manipulations that can easily be detected by participants and that, as with gender, result in the perception of different categories in regards to an attribute (an extravert versus an introvert compared to a more or less extraverted looking person). This approach could be used to ensure that differences within each attribute are comparable across attributes investigated in the study.

All of the suggestions above would shed further light on the impact of differences in facial information on diversity perceptions and choices, but focus more strongly on experimental settings. However, future research could also investigate whether the association between differences in facial information and diversity perceptions and choices could also be found in everyday life. Based on the experimental findings presented here, we would assume that facial differences may impact diversity perceptions in everyday life, when a) other salient attributes such as gender and race are kept constant (see Studies 1a, 1b, 2a, 2b, 3, but also when looking at sports teams, etc.) and people value diversity (Study 4), when b) people are specifically asked to focus on diversity in personality (see, e.g., Study 1b and follow-up Study 5b), and presumably c) when differences in personality matter for the team or judge (see, e.g., Homan & van Kleef, 2021, on diversity on conscientiousness of team members). These conditions could, in many cases, not be met in everyday life.

In general, it is more likely in everyday life that group members will differ on attributes such as gender, race, and age too (as in Study 5). Differences in facial information, furthermore, may be complemented by differences in non-visual information. In a hiring situation (that we mimicked in Studies 3, 4, and 5), recruiters will probably never be asked to make hiring decisions based on profile pictures alone, as they will have access to behavioral information such as skills, experiences, and knowledge of the candidates. Such behavioral information, especially when diagnostic for the role, could override first impressions based on profile pictures (see, e.g., Shen, Mann, & Ferguson, 2020), eventually making diversity in facial information less impactful.

Future research could therefore investigate real life hiring choices for teams to test whether differences in facial information matter. Another pathway could be to research situations in which having a diverse team could be a goal (such as coming up with a list of political candidates for a party, hiring actors for a film cast, or establishing a sounding board at a university or public institution). Here, decision makers will choose between candidates that likely differ on demographic attributes and in perceived personality.

Future research could also test whether choice behavior and the perception of the team can be predicted by demographic attributes alone or whether differences in facial information may explain some additional portion of variance, too. This research would then allow us to answer whether in everyday life differences in facial information may

impact diversity perception and choices, too, or whether the association is only present in certain controlled settings.

12. Conclusion

Based on the research presented in this manuscript, individuals seem to take into account differences in facial information under certain conditions. When keeping other variables such as gender and race constant, individuals consider differences in facial information when asked to evaluate the perceived diversity in teams. Furthermore, they consider differences in facial information when assembling teams, and are more likely to do so when they see value in diversity. This tendency, however, cannot be shown when other variables such as gender differ too. These results contribute to the investigation of perceived diversity and broaden the concept's scope by showing that in certain situations even ambiguous cues such as variety in facial information impacts the perceived diversity of a group, while highlighting boundary conditions when this association is not likely to occur.

Declarations of Competing Interest

None.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sector.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Open practices

The preregistrations for Studies 1b to 5 are available at: (1b) https://aspredicted.org/blind.php?x=/VD7_1DR, (2a) http://aspredicted.org/blind.php?x=br96u2, (2b) https://aspredicted.org/blind.php?x=sk4x5e, (3) https://aspredicted.org/blind.php?x=87bx6e, (4) https://aspredicted.org/blind.php?x=s8bz8d, (5) https://aspredicted.org/blind.php?x=/NMP_AXZ, and (5b, Appendix) https://aspredicted.org/V5Q_Y2S. As a result of the review process and additional studies, the studies are not presented in chronological order.

Acknowledgements

The authors would like to thank Sonja Borner and Helena Brunt for their help in the implementation of the studies. We further thank Laura Fontanesi for her help regarding the programming software for Study 1b and Caroline Tremble for proofreading our manuscript.

Appendix A

A.1. Group picture compilation concept

Group 1: Portrait IDs 01, 03, 18, 22 (Basel Face Database)

	C / $V_{2a}2$ / $V_{2b}11$		A / $V_{2a}12$ / $V_{2b}3$		E / $V_{2a}13$ / $V_{2b}9$		N / $V_{2a}15$ / $V_{2b}10$		O / $V_{2a}6$ / $V_{2b}1$	
	div	non	div	non	div	non	div	non	div	non
01	−	−	−	+	+	−	−	+	−	+
03	−	−	−	+	−	−	+	+	−	+
18	+	−	+	+	+	−	+	+	+	+
22	+	−	+	+	−	−	−	+	+	+
seq	3,1,4,2	3,1,2,4	3,4,1,2	4,3,1,2	4,1,3,2	2,1,4,3	1,3,2,4	4,2,3,1	4,2,1,3	2,3,1,4

Group 2: Portrait IDs 16, 29, 38, 39 (Basel Face Database)

	C / $V_{2a}2$ / $V_{2b}11$		A / $V_{2a}12$ / $V_{2b}3$		E / $V_{2a}13$ / $V_{2b}9$		N / $V_{2a}15$ / $V_{2b}10$		O / $V_{2a}6$ / $V_{2b}1$	
	div	non	div	non	div	non	div	non	div	non
16	−	+	−	+	+	−	+	+	−	−
29	−	+	+	+	−	−	−	+	−	−
38	+	+	−	+	+	−	−	+	+	−
39	+	+	+	+	−	−	+	+	+	−
seq	1,3,4,2	4,2,3,1	1,4,3,2	4,1,3,2	2,3,4,1	4,1,3,2	1,3,4,2	2,4,3,1	4,1,2,3	4,1,3,2

Group 3: Portrait IDs 09,14, 21, 30(Basel Face Database)

	C / $V_{2a}2$ / $V_{2b}11$		A / $V_{2a}12$ / $V_{2b}3$		E / $V_{2a}13$ / $V_{2b}9$		N / $V_{2a}15$ / $V_{2b}10$		O / $V_{2a}6$ / $V_{2b}1$	
	div	non	div	non	div	non	div	non	div	non
09	+	−	+	−	+	+	−	+	+	−
14	+	−	+	−	−	+	−	+	−	−
21	−	−	−	−	−	+	+	+	−	−
30	−	−	−	−	+	+	+	+	+	−
seq	1,3,4,2	1,2,4,3	1,2,4,3	1,4,2,3	3,4,2,1	2,3,4,1	2,4,3,1	4,1,2,3	1,2,3,4	2,1,4,3

Note. C = conscientiousness, A = agreeableness, E = extraversion, N = neuroticism, O = openness; div = diverse manipulation, non = non-diverse / homogeneous manipulation; + = increase in personality trait/vector, − = decrease in personality trait/vector; seq = sequence = the compilation of the portraits in the 2 × 2 grid from upper left to lower right. V = random vector used in Study 2a / Study 2b.

Appendix B

B.1. Further information on the size of the stimuli and background color of all surveys presented in the manuscript

Study	Size of stimuli	Display size in survey	Background color of survey
1a	1360 × 2048, ca. 440 KB	340 × 512	White
1b	1360 × 2048, ca. 440 KB	1360 × 2048	White
2a	1360 × 2048, ca. 440 KB	340 × 512	White
2b	1360 × 2048, ca. 440 KB	340 × 512	White
3	1360 × 2048, ca. 350 KB	340 × 512	White
4	1360 × 2048, ca. 350 KB	340 × 512	White
5	1360 × 2048, ca. 220 KB	340 × 512	White

Appendix C

C.1. Additional mixed model analyses for Study 1a

When computing the trial information variable (ID created out of group number and Big 5 dimension), we were able to create a random effect for trial that had 15 levels. However, one could also model random effects separately for group picture and Big 5. This would more closely capture the nature of our stimuli, but would result in random effects with very few levels only. Bolker (2015) and Westfall, Judd, and Kenny (2014) cautioned against random effects with few levels only due to unstable variance estimates and power concerns. We nevertheless computed a model where we included a fixed effect for diversity, random intercepts for participants, group picture, Big 5, and trial number (order of the trials participants worked on) as well as by-participants, by-group, by-Big 5, and by-trial number random slopes for diversity (= maximal model). We reduced the model complexity step by step until convergence could be reached (Bates, Kliegl, et al., 2015) and omitted correlations between random effects, by-group random slopes for diversity, and by-participants random slopes for diversity. The resulting model does not indicate that the diversity manipulation significantly impacted participants' ratings, $b = 0.14$, $SE\,b = 0.09$, $t(4.49) = 1.48$, $p = .206$ (but keep in mind the problematic power in designs where random effects have few levels only, see Westfall et al., 2014).

To provide a more critical test of whether the impact of the diversity manipulation could be more or less effective when applied to the different Big 5 factors, we computed a linear mixed model for each factor level separately instead of including the variable as a random effect. We included diversity as a fixed effect, a random intercept for participants and a random intercept for group (which is again problematic as this random effect has only three levels, group 1–3). We did not include by-participant or by-group random slopes for diversity to keep models comparable and as including random slopes led to either a singular fit or did not significantly explain variance. Results indicate that the manipulation impacted judgments significantly for agreeableness ($b = 0.21$, $SE\,b = 0.10$, $t(497.00) = 1.98$, $p = .048$) and extraversion ($b = 0.39$, $SE\,b = 0.11$, $t(497.00) = 3.72$, $p < .001$), but not for conscientiousness ($b = 0.04$, $SE\,b = 0.11$, $t(497.00) = 0.36$, $p = .717$), neuroticism ($b = -0.09$, $SE\,b = 0.13$, $t(498.80) = -0.73$, $p = .466$), and openness ($b = 0.17$, $SE\,b = 0.11$, $t(497.00) = 1.57$, $p = .117$). These results tend to corroborate the pattern found when computing the *t*-test (except for neuroticism, see Table 2).

C.2. Additional mixed model analyses for Study 2a

As in Study 1a we additionally estimated the model to include random effects for participants, group pictures, manipulated Big 5 dimension / random vector and trial number. We started with the maximal model and used a stepwise approach to achieve convergence. We estimated a model with diversity, type of manipulation, and their interaction as fixed effects and random intercepts for participants, group picture, and trial number as well as by-Big 5 dimension / random vector, by-group picture, and by-trial number random slopes for diversity and by-group picture random slopes for type of manipulation (the model did not include correlations between random effects). Results, in tendency, corroborated the ANOVA analysis (but keep in mind the problematic power due to few levels of the random effects, see Westfall et al., 2014) and indicated a marginally significant main effect of diversity ($b = 0.13$, $SE\,b = 0.05$, $t(3.80) = 2.39$, $p = .079$), but no significant main effect of type of manipulation ($b = -0.00$, $SE\,b = 0.04$, $t(2.00) = -0.08$, $p = .942$) or of the interaction between the two factors ($b = 0.01$, $SE\,b = 0.09$, $t(7.99) = 0.10$, $p = .921$).

To provide a more critical test of whether the impact of the diversity manipulation could be more or less effective when applied to the different Big 5 or random vectors, we computed a linear mixed model for each factor level separately. As in Study 1a, we computed the simplest model and included diversity as a fixed effect, a random intercept for participants and a random intercept for group. Results indicate that the manipulation impacted judgments significantly for agreeableness ($b = 0.21$, $SE\,b = 0.10$, $t(537.00) = 2.22$, $p = .027$) and extraversion ($b = 0.24$, $SE\,b = 0.10$, $t(537.00) = 2.45$, $p = .015$), but not for openness ($b = 0.20$, $SE\,b = 0.10$, $t(537.00) = 1.96$, $p = .050$), conscientiousness ($b = 0.10$, $SE\,b = 0.10$, $t(537.00) = 0.95$, $p = .342$), and neuroticism ($b = -0.05$, $SE\,b = 0.10$, $t(537.00) = -0.55$, $p = .586$). Turning to the random vectors, results indicate that the manipulation significantly impacted judgments for vector 3 ($b = 0.28$, $SE\,b = 0.10$, $t(537.00) = 2.80$, $p = .005$) and vector 5 ($b = 0.21$, $SE\,b = 0.10$, $t(537.00) = 2.23$, $p = .026$), but not for vector 1 ($b = 0.07$, $SE\,b = 0.10$, $t(537.00) = 0.69$, $p = .492$), vector 2 ($b = -0.13$, $SE\,b = 0.10$, $t(537.00) = -1.28$, $p = .201$), and vector 4 ($b = 0.18$, $SE\,b = 0.10$, $t(537.00) = 1.82$, $p = .069$). Results of these analyses mostly corroborate the *t*-test results (except for the non-significant findings for openness and random vector 4).

C.3. Additional mixed model analyses for Study 2b

As in Studies 1a and 2a we additionally estimated the model to include random effects for participants, group pictures, manipulated Big 5 dimension / random vector and trial number. We started with the maximal model and used a stepwise approach to achieve convergence. We estimated a model with diversity, type of manipulation, and their interaction as fixed effects and random intercepts for participants, group picture, and trial number as well as by-Big 5 dimension / random vector, by-group picture, and by-trial number random slopes for diversity and by-trial number random slopes for type of manipulation (the model did not include correlations between random effects). Results were not significant in this model (but keep in mind the problematic power due to few levels of the random effects, see Westfall et al., 2014), which is true for the main effect of diversity ($b = 0.09$,

SE b = 0.10, *t*(3.17) = 0.92, *p* = .423), the main effect of type of manipulation (*b* = 0.05, *SE b* = 0.03, *t*(56.79) = 1.59, *p* = .118) and the interaction between the two factors (*b* = 0.02, *SE b* = 0.11, *t*(8.01) = 0.14, *p* = .891).

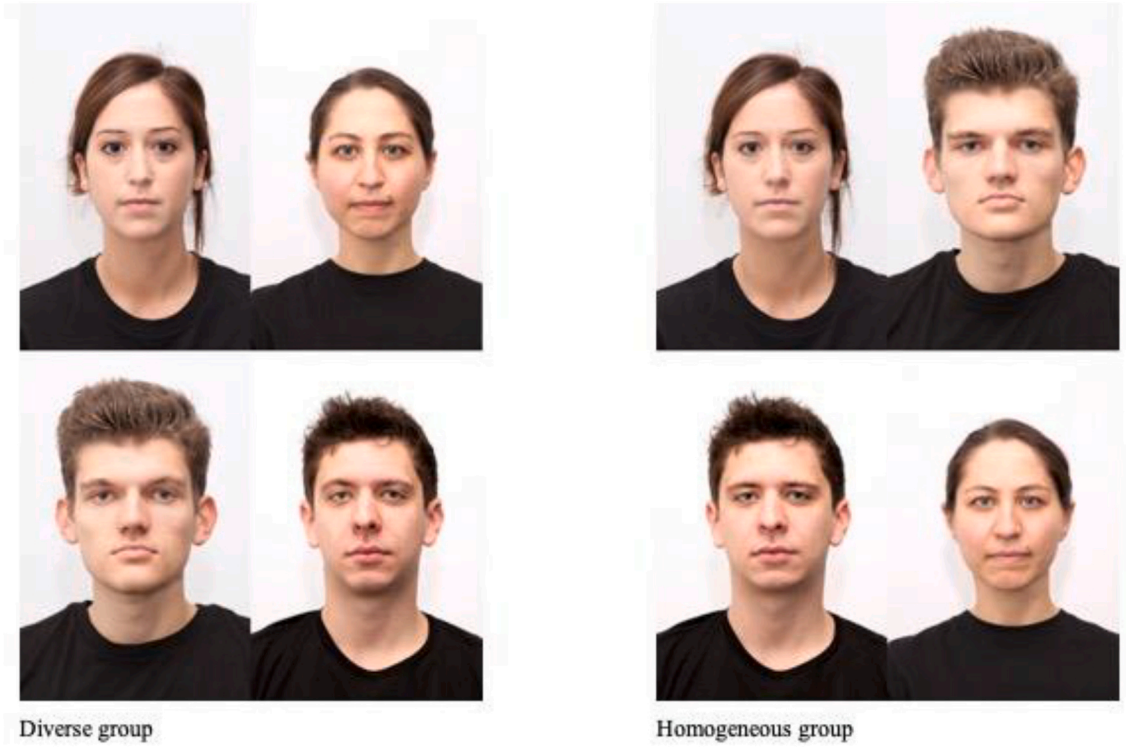
To provide a more critical test of whether the impact of the diversity manipulation could be more or less effective when applied to the different Big 5 or random vectors, we computed a linear mixed model for each factor level separately. As in Study 1a and 2a, we computed the simplest model and included diversity as a fixed effect, a random intercept for participants and a random intercept for group. Results indicate that the manipulation impacted judgments significantly for agreeableness (*b* = 0.25, *SE b* = 0.11, *t*(517.00) = 2.35, *p* = .019) and extraversion (*b* = 0.32, *SE b* = 0.10, *t*(517.00) = 3.06, *p* = .002), but not for openness (*b* = 0.13, *SE b* = 0.10, *t*(517.00) = 1.30, *p* = .193), conscientiousness (*b* = −0.10, *SE b* = 0.10, *t*(517.00) = −1.00, *p* = .318), and neuroticism (*b* = −0.13, *SE b* = 0.10, *t*(517.00) = −1.22, *p* = .223). Turning to the random vectors, results indicate that the manipulation impacted judgments significantly for vector 3 (*b* = 0.29, *SE b* = 0.10, *t*(517.00) = 2.92, *p* = .004), but not for vector 1 (*b* = 0.02, *SE b* = 0.10, *t*(517.00) = 0.16, *p* = .870), vector 2 (*b* = −0.05, *SE b* = 0.09, *t*(517.00) = −0.55, *p* = .586), vector 4 (*b* = 0.17, *SE b* = 0.10, *t*(517.00) = 1.74, *p* = .082), and vector 5 (*b* = 0.00, *SE b* = 0.10, *t*(517.00) = 0.00, *p* = .975). Results of these analyses mostly corroborate the *t*-test results (except for the non-significant findings for random vector 4).

C.4. Additional analyses for Study 5

When interpreting the results reported in the manuscript, we identified a confound in the exploratory analyses. Participants being more likely to assemble diverse teams in Type 2-aligned trials (see preregistered analyses) and the association between diversity influence and choices being stronger in Type 2-aligned versus Type 1-misaligned trials (exploratory analyses) could be driven by the choice and focus on female as well as facially different candidates (as in Type 2-aligned trials, the facially different candidate was also the female candidate, meaning that personality and demographic associated diversity were aligned). To better understand the impact of facial variation more generally, we recoded the dependent variable so that 1 indicates choice of female candidate and 0 choice of the male candidate. We computed a model that included type of trial, diversity influence, and their interaction as fixed effects, random intercepts for participants and trials, and by-participants random slopes for type of trial. Results indicate a significant intercept (*b* = 0.56, *SE b* = 0.12, *z* = 4.75, *p* < .001), indicating that choice of the female candidate was above chance level (participants chose the female candidate in on average 85.45 out of 144 trials). The effect of type of trial was not significant, *b* = 0.02, *SE b* = 0.11, *z* = 0.17, *p* = .862. Diversity influence was significantly and positively associated with choosing the female candidate, *b* = 0.29, *SE b* = 0.12, *z* = 2.52, *p* = .012, and this association differed for type of trial levels, as indicated by a significant interaction effect, *b* = 0.26, *SE b* = 0.10, *z* = 2.57, *p* = .010. To disentangle the latter, we computed two separate models for each type of trial factor level. Diversity influence was included as a fixed effect and we included random intercepts for participants and trials. Results indicated a positive but non-significant association between diversity influence and choice of female candidates in Type 1-misaligned trials, *b* = 0.14, *SE b* = 0.12, *z* = 1.14, *p* = .253, and a positive association in Type 2-aligned trials, *b* = 0.40, *SE b* = 0.11, *z* = 3.60, *p* < .001. The stronger the self-indicated influence of diversity values, the more likely participants were to choose female candidates – and this tendency was exacerbated in trials where the female candidate was also facially different in terms of extraversion from the existing two-person team (Type 2-aligned trials).

Appendix D

D.1. Illustration of random vector 3 from Study 2a (V12)



Individual manipulations (on the left side the manipulation is reduced and on the right side the manipulation is enhanced):

Appendix E

E.1. Follow-up Study 5b

Study 5 may be considered a conservative test to investigate whether differences in facial information may impact diversity choices when groups differ on other, more salient variables (e.g., gender) too. When instructing participants to focus on diversity, they may be immediately more likely to focus on identity-relevant attributes such as gender, as differences here are more strongly associated with the concept of diversity.

We therefore conducted a follow-up study, which was almost identical to Study 5: Participants were again asked to choose candidates for a two-person team and were informed that when doing so they should keep in mind the company's core value of diversity. Different from the original Study 5, however, we introduced a systematic variation in the instructions: Half of the participants were explicitly asked to focus on personality diversity and the other half on gender diversity. Participants then worked on 48 trials (instead of 144 to make the study length comparable to the previous Studies 2 and 3), asked for demographic information, and thanked for their participation. Different from Study 5 we did not ask them to what extent diversity impacted their decisions. The follow-up study was preregistered, https://aspredicted.org/V5Q_Y2S, and the data of 200 UK participants were collected via Prolific.

One hundred and ninety-nine individuals started the study and provided informed consent. In line with preregistered exclusion criteria, we excluded participants due to low carefulness ratings ($n = 6$) and when they indicated that images sometimes or always did not appear ($n = 1$). The resulting sample consisted of 192 participants ($M_{age} = 38.57$, $SD = 13.41$; 79 male, 111 female, 2 non-binary).

When investigating personality diversity by coding the choice of the facially diverse candidate (in relation to the two-person team) as 1 and the facially homogeneous candidate as 0, participants on average selected facial diversity in 25.95 out of 48 trials ($SD = 5.84$). This number was higher when participants were instructed to focus on personality diversity, $M = 27.71$, $SD = 6.77$, and lower when participants were instructed to focus on gender diversity, $M = 24.23$, $SD = 4.10$.

We then used a generalized linear mixed model for binary outcomes and included condition ($-0.5 =$ *gender diversity*; $0.5 =$ *personality diversity*) and type of trial ($-0.5 =$ *Type 1-misaligned trials*; $0.5 =$ *Type 2-aligned trials*) as fixed effects in the model. As random effects we included random intercepts for participant and trials and by-participant random slopes for type of trial and by-trials random sloped for condition. Choice of the facially diverse (coded as 1) versus homogeneous team (coded as 0) served as the dependent variable. Results indicate a significant intercept, $b = 0.20$, $SE\ b = 0.05$, $z = 4.17$, $p < .001$, showing that participants' choice of facially diverse candidates was higher than chance level. We further found a significant effect of condition, $b = 0.33$, $SE\ b = 0.09$, $z = 3.50$, $p < .001$, indicating that participants with the personality diversity instructions (vs. gender diversity instructions) were more likely to choose the facially diverse team. Results further yield a significant main effect of type of trial, $b = 1.15$, $SE\ b = 0.16$, $z = 7.37$, $p < .001$, showing that choice of a facially diverse candidate was more likely in Type 2-aligned trials compared to Type 1-misaligned trials.

When focusing on gender diversity by coding the choice of the female candidate as 1 and the choice of the male candidate as 0, participants on average selected gender diversity in 28.94 out of 48 trials ($SD = 8.29$). This number was higher when participants were instructed to focus on gender diversity, $M = 32.20$, $SD = 8.69$, and lower when participants were instructed to focus on personality diversity, $M = 25.61$, $SD = 6.35$.

We then used a generalized linear mixed model for binary outcomes and included condition ($-0.5 =$ *gender diversity*; $0.5 =$ *personality diversity*) and type of trial ($-0.5 =$ *Type 1-misaligned trials*; $0.5 =$ *Type 2-aligned trials*) as fixed effects in the model. As random effects we included random intercepts for participant and trials and by-participant random slopes for type of trial and by-trials random sloped for condition. Choice of the gender diverse (coded as 1) versus homogeneous team (coded as 0) served as the dependent variable. Results indicate a significant intercept, $b = 0.55$, $SE\ b = 0.07$, $z = 7.56$, $p < .001$, showing that participants' choice of female candidates was higher than chance level. We further found a significant effect of condition, $b = -0.75$, $SE\ b = 0.15$, $z = -5.10$, $p < .001$, indicating that participants with the gender diversity instructions (vs. personality diversity instructions) were more likely to choose the team with the female candidate. Results further yield a significant main effect of type of trial, $b = 0.38$, $SE\ b = 0.10$, $z = 3.76$, $p < .001$, showing that choice of a female candidate was more likely in Type 2-aligned trials compared to Type 1-misaligned trials.

References

Alt, N. P., Goodale, B., Lick, D. J., & Johnson, K. L. (2019). Threat in the company of men: Ensemble perception and threat evaluations of groups varying in sex ratio. *Social Psychological and Personality Science, 10*(2), 152–159. https://doi.org/10.1177/1948550617731498

Alt, N. P., & Phillips, L. T. (2021). Person perception, meet people perception: Exploring the social vision of groups. *Perspectives on Psychological Science, 1–20*. https://doi.org/10.1177/17456916211017858

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. http://arxiv.org/abs/1506.04967.

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

Bolker, B. M. (2015). Linear and generalized linear mixed models. In G. A. Fox, S. Negrete-Yankelevich, & V. J. Sosa (Eds.), *Ecological statistics: Contemporary theory and application* (pp. 309–333). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199672547.001.0001.

Cunningham, G. B. (2007). Perceptions as reality: The influence of actual and perceived demographic dissimilarity. *Journal of Business Psychology, 22*, 79–89. https://doi.org/10.1007/s10869-007-9052-y

Daniels, D. P., Neale, M. A., & Greer, L. L. (2017). Spillover bias in diversity judgment. *Organizational Behavior and Human Decision Processes, 139*, 92–105. https://doi.org/10.1016/j.obhdp.2016.12.005

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*, 175–191. https://doi.org/10.3758/BF03193146

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review, 118*(2), 247–279. https://doi.org/10.1037/a0022327

Haberman, J., Lee, P., & Whitney, D. (2015). Mixed emotions: Sensitivity to facial variance in a crowd of faces. *Journal of Vision, 15*(4), 1–11. https://doi.org/10.1167/15.4.16

Harrison, D. A., & Klein, K. J. (2007). What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Review, 32*(4), 1199–1228. https://doi.org/10.5465/AMR.2007.26586096

Harrison, D. A., Price, K. H., & Bell, M. P. (1998). Beyond relational demography: Time and the effects of surface- and deep-level diversity on work group cohesion. *Academy of Management Journal, 41*(1), 96–107. https://doi.org/10.2307/256901

Harrison, D. A., Price, K. H., Gavin, J. H., & Florey, A. T. (2002). Time, teams, and task performance: Changing effects of surface- and deep-level diversity on group functioning. *Academy of Management Journal, 45*(5), 1029–1045. https://doi.org/10.2307/3069328

Hentschel, T., Shemla, M., Wegge, J., & Kearney, E. (2013). Perceived diversity and team functioning: The role of diversity beliefs and affect. *Small Group Research, 44*(1), 33–61. https://doi.org/10.1177/1046496412470725

Hobman, E. V., Bordia, P., & Gallois, C. (2003). Consequences of feeling dissimilar from others in a work team. *Journal of Business and Psychology, 17*, 301–325. https://doi.org/10.1023/A:1022837207241

Homan, A. C., Greer, L. L., Jehn, K. A., & Koning, L. (2010). Believing shapes seeing: The impact of diversity beliefs on the construal of group composition. *Group Processes & Intergroup Relations, 13*(4), 477–493. https://doi.org/10.1177/1368430209350747

Homan, A. C., & van Kleef, G. A. (2021). Managing team conscientiousness diversity: The role of leader emotion-regulation knowledge. *Small Group Research, 1–31*. https://doi.org/10.1177/10464964211045015

Homan, A. C., van Knippenberg, D., Van Kleef, G. A., & De Dreu, C. K. W. (2007). Bridging faultlines by valuing diversity: Diversity beliefs, information elaboration, and performance in diverse work groups. *Journal of Applied Psychology, 92*(5), 1189–1199. https://doi.org/10.1037/0021-9010.92.5.1189

Jaeger, B., & Jones, A. L. (2021). *Which facial features are central in impression formation* (pp. 1–23). https://doi.org/10.31234/osf.io/9c57t

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin, & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). Guilford Press.

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. https://doi.org/10.1037/a0028347

Klapper, A., Dotsch, R., van Rooij, I., & Wigboldus, D. H. J. (2016). Do we spontaneously form stable trustworthiness impressions from facial appearance? *Journal of Personality and Social Psychology, 111*(5), 655–664. https://doi.org/10.1037/pspa0000062

van Knippenberg, D., & Haslam, S. A. (2003). Realizing the diversity dividend: Exploring the subtle interplay between identity, ideology, and reality. In S. A. Haslam, D. van Knippenberg, M. J. Platow, & N. Ellemers (Eds.), *Social identity at work: Developing theory for organizational practice* (pp. 61–77). Psychology Press.

van Knippenberg, D., Haslam, S. A., & Platow, M. J. (2007). Unity through diversity: Value-in-diversity beliefs, work group diversity, and group identification. *Group Dynamics: Theory, Research, and Practice, 11*(3), 207–222. https://doi.org/10.1037/1089-2699.11.3.207

Kubota, J. T., & Ito, T. A. (2007). Multiple cues in social perception: The time course of processing race and facial expression. *Journal of Experimental Social Psychology, 43*, 738–752. https://doi.org/10.1016/j.jesp.2006.10.023

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist, 52*(5), 509–516. https://doi.org/10.1037/0003-066x.52.5.509

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. https://doi.org/10.1016/j.jesp.2009.12.002

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*(32), 11087–11092. https://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 *Advanced video and signal based surveillance, 2009. AVSS'09. Sixth IEEE international conference* (pp. 296–301). https://doi.org/10.1109/AVSS.2009.58

Phillips, L. T., Slepian, M. L., & Hughes, B. L. (2018). Perceiving groups: The people perception of diversity and hierarchy. *Journal of Personality and Social Psychology, 114*(5), 766–785. https://doi.org/10.1037/pspi0000120

Phillips, L. T., Weisbuch, M., & Ambady, N. (2014). People perception: Social vision of groups and consequences for organizing and interacting. *Research in Organizational Behavior, 34*, 101–127. https://doi.org/10.1016/j.riob.2014.10.001

Prolific Academic. (2018). Prolific Academic. https://prolific.ac/.

Rudert, S. C., Keller, M. D., Hales, A. H., Walker, M., & Greifeneder, R. (2020). Who gets ostracized? A personality perspective on risk and protective factors of ostracism. *Journal of Personality and Social Psychology, 18*(6), 1247–1268. https://doi.org/10.1037/pspp0000271

Shemla, M., & Meyer, B. (2012). Bridging diversity in organizations and cross-cultural work psychology by studying perceived differences. *Industrial and Organizational Psychology, 5*(3), 370–372. https://doi.org/10.1111/j.1754-9434.2012.01464.x

Shemla, M., Meyer, B., Greer, L., & Jehn, K. A. (2016). A review of perceived diversity in teams: Does how members perceive their team's composition affect team processes and outcomes? *Journal of Organizational Behavior, 37*(1), 89–106. https://doi.org/10.1002/job.1957

Shen, X., Mann, T. C., & Ferguson, M. J. (2020). Beware a dishonest face?: Updating face-based implicit impressions using diagnostic behavioral information. *Journal of Experimental Social Psychology, 86*(July 2019), 103888. https://doi.org/10.1016/j.jesp.2019.103888

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*(37), 9210–9215. https://doi.org/10.1073/pnas.1807222115

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition, 127*(1), 105–118. https://doi.org/10.1016/j.cognition.2012.12.001

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

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27*(6), 813–833. https://doi.org/10.1521/soco.2009.27.6.813

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

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. https://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), Article e0193190. https://doi.org/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. https://doi.org/10.1037/pspp0000064

Westfall, J. (2016). *PANGEA: Power ANalysis for GEneral anova designs. Working paper* (pp. 1–33). http://jakewestfall.org/publications/pangea.pdf.

Westfall, J., Judd, C. M., & Kenny, D. A. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Williams, K. Y., & O'Reilly, C. A. I. (1998). Demography and diversity in organizations: A review of 40 years of research. In B. M. Staw, & L. L. Cummings (Eds.), *Vol. 20. Research in organizational behavior* (pp. 77–140). JAI Press.

Willis, J., & Todorov, A. (2016). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass, 2*(3), 1497–1517. https://doi.org/10.1111/j.1751-9004.2008.00109.x