Neural Correlates of Social Group Perception – Mapping Perceived Social Categories in Brain Activation

Doctoral thesis

to obtain a doctorate (PhD)

from the Faculty of Medicine

of the University of Bonn

Omar Salah Ahmed

from Dakahlia, Egypt

2025

Written with authorization of

the Faculty of Medicine of the University of Bonn

First reviewer: PD. Dr. Johannes Schultz

Second reviewer: Prof. Dr. Silke Lux

Day of oral examination: 13th of November 2025

From the Center for Economics and Neuroscience (CENs), Institut für experimentelle Epileptologie und Kognitionsforschung (IEECR)

Table of contents:

List of abbreviations:

ACC	Anterior Cingulate Cortex
AH:	Arts and Humanities
ATL	Anterior Temporal Lobe
BG	Between-group similarity
BGD	Between-group discriminability
BIDS	Brain Imaging Data Structure
BOLD	Blood Oxygen Level Dependent
df	Degrees of freedom
DLPFC	Dorsolateral Prefrontal Cortex
EEG	Electroencephalography
EPI	Echo Planar Imaging
FFA	Fusiform Face Area
fMRI	Functional Magnetic Resonance Imaging
fNIRS	Functional Near-Infrared spectroscopy
GLM	General Linear Model
GRAPPA	Generalized Auto-calibrating Partially Parallel Acquisitions
HRF	Hemodynamic Response Function
LDC	Linear Discriminant Contrast
LDt	Linear Discriminant t-value
MDS	Multi-dimensional Scaling
MPRAGE	Magnetization-Prepared Rapid Gradient Echo Imaging
MVPA	Multivariate Pattern Analysis
NAT:	Natural Sciences
ODT:	Optimal Distinctiveness Theory
PANAS	Positive and Negative Affect Schedule
PER:	Percentage Estimate Task
RDM	Representational Dissimilarity Matrix
RGM	Rating Group Members Task
ROI	Region of Interest
RSA	Representational Similarity Analysis
SAQ:	Self-assessment Questionnaire
SIM:	Similarity Task
STG	Superior Temporal Gyrus
STS	Superior Temporal Sulcus
TFCE	Threshold-free Cluster Enhancement
WG	Within-group similarity
WGD	Within-group Discriminability

1 Introduction

According to social identity theory, a person's social identity depends on the similarities and differences (e.g., in locality, race, or language) between that person and other people. Based on these similarities and differences, a person categorizes herself and similar others as in-group members and different others as out-group members (Turner, 2010). This categorization is thought to contribute to the maintenance and enhancement of social identity (Abrams & Hogg, 1990; Tajfel & Turner, 1979), which is thought to be critical for a successful group life (Lau & Cikara, 2017). However, such categorizations can lead to social effects such as in-group bias, stereotyping, dehumanizing outgroup members, (Bruneau, 2018; Harris & Fiske, 2006; Haslam & Stratemeyer, 2016; Kersbergen & Robinson, 2019), and perceived outgroup threat (Lantos et al., 2020; Pickett & Brewer, 2001; Riek et al., 2006; Turner et al., 1987; Wilson & Hugenberg, 2010), which in turn facilitates different positive or negative behaviors towards the individual, like discrimination, aggression (Rai et al., 2017; Workman et al., 2020), or empathy (Han, 2018; Ruckmann et al., 2015; Vanman, 2016).

These behavioral effects are likely to be related to neural representations of the people one interacts with, specifically the impact of an individual's group membership on their neural representation. Previous studies have revealed the impact of group membership, for example, on the neural responses evoked when observing people interacting socially, (Katsumi & Dolcos, 2018; Molenberghs et al., 2016) or experiencing pain (Shen et al., 2018). Faces of people from one's own social group were shown to evoke different neural responses compared to faces of another social group, whether the groups were based on race, (Cunningham et al., 2004; Farmer et al., 2020; Hart et al., 2000, p. 202; Phelps et al., 2000) or arbitrary cues (Contreras-Huerta et al., 2014; Krautheim et al., 2018; Van Bavel et al., 2008). However, the neural correlates of the perceived similarities and differences between members of a person's own and other social groups, a key component of social categorization, are still unknown.

1.1 Aim of the Thesis

The three experiments we conducted for this thesis aimed to: a) Test previously established group-based measures of perceived group homogeneity and compare them with individual-based measures. b) Establish a pairwise-comparison-based behavioral

measure of perceived similarity between social group members, in order to directly compare similarities in perception with similarities in brain activation. c) Find which brain regions reflect changes in perceived similarities induced by social categorization are involved in changing the perception of the similarity between newly encountered persons. d) Find which brain regions show activation patterns that reflect the behaviorally recoded perceived similarities between persons.

2 Study 1: Replicating Previous Measures of Group Homogeneity

2.1 Introduction

Since the 1950s, there has been an interest in quantifying perceived group homogeneity and studying the psychological framework underlying stereotyping and discriminatory actions against outgroups. (G. M. Gilbert, 1951; Karlins et al., 1969; Katz & Braly, 1933) One of the theoretical frameworks, called Optimal Distinctiveness Theory (ODT), originated from Marilyn Brewer's work in the 1990s and postulates that individuals oscillate between two primary drives. The first is the *Assimilation need,* and the second is the *Differentiation need* (Brewer, 1991). Assimilation refers to the motivation to be included in a social group, while differentiation is the desire to have clearly differentiated social groups. The theory builds on the social identity theory, and the experiments done by Brewer show that the level of assimilation to a group and the level of group differentiation can affect the social identity (Brewer & Pickett, 1999; Pickett & Brewer, 2001).

2.1.1 Measures of Group Homogeneity

There are multiple measures of group homogeneity with variable consistency. Ostrom & Sedikides (1992) discussed seven tasks that have been used since the inception of social categorization theory in the 1970s. Although the meta-analysis demonstrated that outgroup homogeneity was measurable across different traits, social groups, and experimental paradigms, it also highlighted the variability in results with different tasks and the nature of the social group, whether it was a *Natural Group* or a *Minimal Group* (Ostrom & Sedikides, 1992). In our first experiment, we attempted to compare two previously established tasks, namely the Percentage Estimate Task (PER) and the Similarity task (SIM), which are based on group-level judgments, with a new task, the Rating Group Members Task (RGM), which is based on individual-level judgments. The similarity task directly measures group homogeneity, where groups with high perceived similarity among their members are considered homogeneous. While PER task uses stereotypes as a proxy for homogeneity, where groups with a majority of members hold stereotypes typical of this group and only a few members hold nontypical stereotypes (for more information about SIM and PER, see Boldry et al. (2007). An example of typical stereotypes could be being "logical" for Natural science students or being "creative" for Arts and Humanities students.

2.1.2 Trait Relevance and Typicality

It is essential to note that since social categorization is a result of similarities and differences in certain traits, the nature of the social category itself determines which personal traits are relevant to it and which are irrelevant. According to the prototype theory by Rosch & Lloyd, 1978, each social category should have central (or prototypical) members with certain diagnostic traits that perfectly describe the social group. The diagnostic traits are then “essential” and “relevant” to the social group. Any individual who is categorized as belonging to a particular social group must possess these diagnostic traits to a certain extent. Although Medin & Schaffer suggested that people categorize based on stored exemplars rather than prototypes (central examples) in their Exemplar theory (Medin & Schaffer, 1978), both theoretical frameworks agreed that for any social category, there are relevant traits that define the members who belong to the category. Therefore, the social categorization of an individual is influenced more by “relevant” traits than by “irrelevant” traits (E. R. Smith & Zarate, 1990; E. R. Smith & Zárate, 1992).

Applying this to the categorization of students, we can claim that in our case (during the summer of 2022 in Bonn, Germany, where the political atmosphere is not overly charged), the political orientation of the students is less relevant than how much a student is “logical” and “evidence-oriented” in inferring the group membership of the student. As we will show, natural science students were perceived to be more logical than Arts and Humanities students. However, we cannot assume that trait relevance is binary, as some seemingly irrelevant traits could still be perceived to be indirectly relevant. For example, being a Natural science student might affect a student’s lifestyle, eventually resulting in the tendency of natural science students to be overweight. In this case, it would be easier to infer that an overweight student is a Natural science student in comparison to a lean student.

2.1.3 Effect of the Grouping Paradigm

Natural groups like Study fields have some inherited problems that experimenters cannot control. Some confounders, such as participants’ prior knowledge about the group or their familiarity with group members, can influence the perception of group members unexpectedly. For instance, if we ask an Arts and Humanities student to rate the perceived

group homogeneity of natural science students, a previous interaction—either positive or negative—with natural science students could bias the rating.

An alternative option to using a natural group like “Study Field” is to use a minimal group paradigm. This is where experimenters create new experimental social groups in the lab and assign participants randomly to one of the groups. The minimal group paradigm was first introduced by (Tajfel et al., 1971) and has since then been used in social experiments to measure perceived group homogeneity. It was also used in fMRI experiments to compare neural activation between ingroups and outgroups, as it offers several advantages over natural groups. Minimal group paradigms provide a more controlled environment for testing social group perception and intergroup interactions, since minimal groups are devoid of associated stereotypes. The minimal group paradigm can be used to construct a “mixed group paradigm,” in which each minimal group comprises members from multiple natural groups. The mixed group paradigm has been used recently in neuroscience experiments to distinguish effects related to natural groups from effects associated with social categorization into ingroups and outgroups (Cao et al., 2015; Feng et al., 2011; Holyoak & Morrison, 2012; Van Bavel et al., 2008).

Unfortunately, minimal group paradigms are more challenging to implement, as they require training of participants to remember the group membership of each person, as in Van Bavel et al., 2008. In addition, a meta-analysis conducted by Mullen & Hu, (1989) and a review by Ostrom & Sedikides, (1992) concluded that the outgroup-homogeneity effect is less evident in experiments that used minimal group paradigms. For these reasons, we opted to use a natural group paradigm.

2.1.4 Social Categorization Versus Stereotype Application

When discussing social categories and stereotyping, it is essential to distinguish between two distinct processes. The first one involves categorizing an individual into a social group based on certain traits observed in that individual. The second is stereotype application, in which people first know the group membership of a specific individual, and then they start associating a previously established stereotype about the social group with the individual. This distinction between the two processes was proposed as dual-process models of person perception, which argue that both bottom-up (trait-driven) and top-down

(social category-driven) processes contribute to social judgments (Bin Meshar et al., 2021; Freeman & Ambady, 2011; Stolier & Freeman, 2017).

In social categorization (the bottom-up process), people would infer the social category of the individual after learning the different traits of this person. For example, in an experimental paradigm, when showing a video of an individual to participants where the subject doesn't explicitly mention their membership to a specific social group but instead shows some specific physical features (e.g. specific tattoos, head veil, overweight, etc.) or behavioral traits (e.g. eco-awareness, interest in sports, creativity, etc.) participants could infer certain social categories of the individual. An example of an experimental paradigm that triggers the top-down process of stereotype application would be one in which participants see a video of an individual who explicitly says that they belong to the "physics students" group. Stereotype application would be assuming that the individual also possesses other traits prototypical of "physics students". In short, social categorization starts by learning traits of an individual, while stereotype application starts by knowing the group membership of the individual. It's not uncommon that both processes happen simultaneously while the participants learn more about the individual.

2.1.5 Social Category Salience

Individuals are naturally categorized into multiple social groups (Crisp et al., 2010), for example, a Natural science student could also belong to the groups of "young people", "educated people", "environment-friendly people", "football team X supporters", etc. Although any individual is a member of tens of social groups, only a few social categories are salient at a time (Dovidio et al., 2006). This depends on the context, for example, in a football match, the social category of being "football team X supporter" is more salient, while in a political discussion, belonging to the group of "environment-friendly people" or "educated people" becomes more salient than other group memberships. This is why, in an experimental paradigm, it's important to control which social group membership is salient.

2.2 Study Aim

This study aimed to replicate the experiment done by Pickett & Brewer (2001) in which they measured the influence of assimilation and differentiation needs on the perceived ingroup and outgroup homogeneity. More specifically, we wanted to replicate the influence

of inducing *social identity threat* on perceived out-group homogeneity. Additionally, our second goal was to test and compare various measures of perceived group homogeneity. We aimed to determine whether group-level judgments are representative of judgments of individuals who belong to the same group. Because our participants were not familiar with the members of the ingroup and outgroup, we didn't expect that individual-specific information about group members would bias the individual-based ratings of perceived group homogeneity or stereotypicality judgments of members.

2.3 Hypotheses

2.3.1 Hypothesis 1

Ratings of the similarity in personality and social skills, and ratings of stereotypicality, will be significantly higher for the outgroup than for the ingroup.

2.3.2 Hypothesis 2

The stereotypicality ratings of a group will be positively correlated with the stereotypicality ratings of individual group members.

2.3.3 Hypothesis 3

An induction of a social identity threat will result in significantly higher perceived group homogeneity for both ingroup and outgroup compared to control conditions.

More specifically, we divided the third hypothesis into two parts. First, we tried to reproduce the analysis done by Pickett & Brewer (2001) to point out possible differences. The hypothesis 3.1 was thus:

Participants who received a social identity threat intervention will rate the ingroup and outgroup higher in terms of similarity and percentage estimates on stereotype-relevant and -irrelevant traits than participants of the control condition (no threat).

The second part of the third hypothesis involved the adapted study design to prepare for the fMRI study. We used a pre vs. post study design to account for interpersonal differences in ratings. To include this design in the analysis, our hypothesis 3.2 was:

Participants who received a social identity threat intervention will have a bigger change in ratings of similarity and stereotypicality from before to after the intervention than participants in the control condition (no threat).

2.4 Study Design

The study used a between-subjects design. We recruited Natural Science students (NAT) and Arts and Humanities students (AH) and asked them to fill out an irrelevant questionnaire supposedly measuring some self-attributes. At the same time point, we asked them to rate the perceived ingroup and outgroup homogeneity using the group-level tasks from Pickett & Brewer's study (stereotypicality and similarity tasks) and our new individual-based task. The participants were split into the *Identity threat group* and the *Control group.* Participants in the threat group were told that their score in the questionnaire was very different from the scores of their ingroup while in the control group participants were told that their scores were similar to their ingroup. In both conditions, participants conducted the three tasks before and after the identity threat intervention.

2.5 Pretesting

We decided to replicate Pickett & Brewer (2001) as much as possible, so we started by replicating the pretesting survey among bachelor students. We conducted the pretesting online in German for Bachelor students all over Germany. Firstly, we wanted to ensure that participants from both groups perceived students of Arts and Humanities and students of Natural Sciences as two distinct groups. Secondly, in the pretesting questionnaire, participants reported which three study fields represented each group.

2.5.1 Questionnaire

In the questionnaire, we asked participants how strongly they identify as either Arts and Humanities (AH) or Natural Science (NAT) students. Then they selected from a list of 11 NAT-related and 28 AH-related study topics the top three study topics that they considered stereotypical of each study field. We also asked them to indicate from a list of 100 attributes from academic, personal, social, and general aspects, for example, "lernbegierig", "kreativ" or "gierig" the most stereotypical characteristics associated with each study field.

2.5.2 Participants

We collected data online from 36 bachelor students all over Germany.

2.5.3 Results

We selected the top three study topics that were chosen by at least half of the participants i.e. 18 participants to be stereotypical of the study field. We used a cutoff of 50%, meaning

that we only considered study topics and traits that were selected by more than half of the participants.

2.5.3.1 Stereotypical Study Topics

The two most typically rated study fields for each group were used for the recruitment of participants for the video recordings that were used as the experiment stimuli (illustrated below).

2.5.3.2 Stereotypical Traits

The top four traits that were chosen by more than 50 % of the participants for each group were considered stereotype-relevant traits for that group. Stereotype-relevant traits for each group simultaneously served as counter-stereotypic traits for the respective other group. There was no overlap in the chosen traits for the groups, ensuring the distinctive difference in prototypic perception of the groups and consensus in terms of the group stereotypes. Traits that were chosen zero times by neither of the groups were considered stereotype-irrelevant traits for both groups.

Table 1: List of typical study fields and stereotype-relevant and -irrelevant traits for AH and NAT study fields as assessed in the Pretesting Study. Typical study topics and stereotype-relevant traits were chosen by more than 50% of the participants. Stereotype-irrelevant traits were chosen zero times by neither of the groups.

Study field	Arts and Humanities	Natural science
Typical study topics	Politics	Chemistry
	History	Physics
	Philosophy	Biology
Stereotype-relevant traits	Creative Social Expressive (Wortgewandt) Sociable (Kontaktfreudig)	Disciplined Logical Analytical eager to learn
Stereotype-irrelevant traits	Greedy Passive	

2.6 Materials & Methods

2.6.1 Stimuli

2.6.1.1 Introductory Videos

Eleven German bachelor students recorded short videos of themselves. In each video, each of the eight students depicted in our stimuli mentioned their study field, and which group they categorized themselves into i.e. either “Natural Science” or “Arts and Humanities” and then explained why they considered themselves as members of this social group. They were required to identify themselves with the respective group and to sign a consent form and data protection agreement. The recordings were held on the videoconferencing platform Zoom (Zoom Inc.). The participants were required to wear a monochrome t-shirt, preferably in white or grey. The participants were asked to place themselves in front of a white wall and to hold their device such that their face was positioned in the middle of the screen. The participants were not wearing, head or face accessories, they had no visible tattoos or any symbol representative of other social groups (e.g., religious marks). The videos were motion stabilized.

2.6.1.2 Color Portraits

We took a color photo of each of the eleven students before or after recording the video. The participants were asked to hold a neutral facial expression. The portraits showed the students’ faces, part of their T-shirts, and the white background.

2.6.2 Social Groups

We opted for a natural group, which we called “Study Field” which could be either “Natural Science” (NAT) or “Arts and Humanities” (AH). These two social groups played a key role in our experiment. We recruited participants who only belonged to either one of the social groups. Also, our stimuli and the experimental tasks focused on rating members of these social groups.

2.6.3 Experimental Tasks:

2.6.3.1 Self-Assessment Questionnaire (SAQ)

The SAQ consisted of several attributes on which participants were asked to rate themselves (e.g., physical attractiveness, discipline, leadership abilities). Participants were also asked to rate the level of confidence in their ratings of these attributes. (A copy

of the SAQ is included in Appendix 11.1) Participants were led to believe that their score in the SAQ task provided important information about their insight into their capabilities. Since SAQ requires participants to rate their abilities relative to those of other students, this context would increase the attention paid to the relative differences between themselves and other students, allowing for comparison between their scores and those of their ingroup and outgroup. This, in turn, would influence the perceived social identity and perceived social group homogeneity. We intentionally wanted to confuse the participants about how SAQ scores were calculated, so that they couldn't guess their actual scores and would not lose confidence in the results we provided.

2.6.3.2 Similarity Task Questionnaire (SIM)

This questionnaire, which was used before in some experiments, e.g., in Pickett & Brewer 2001 is based on group-level judgement of social groups. In the questionnaire, participants are asked to rate on a 7-point Likert scale the similarity between all the members of a certain group in a certain trait or a dimension. In our case, we asked about the similarity of group members to each other in a) Personality and b) Social skills. The similarity task is a direct and explicit measure of the similarity of group members across dimensions such as personality or social life (Park & Judd, 1990). The exact questions we asked were:
"Wie ähnlich denken Sie sind Geisteswissenschaftsstudent*innen / Naturwissenschaftsstudent*innen bezüglich Ihrer Persönlichkeit / sozialen Fähigkeiten zueinander?"

2.6.3.3 Percentage Estimate Task (PER)

From the pretesting, we formulated a list of 10 traits, four of which were stereotypical of NAT students, four were stereotypical of AH students, and two were not stereotypical of any of the groups. Park and Judd (1990) and later researchers considered that rating a group to have a high percentage of stereotypical traits, for example being analytical or seeking knowledge for Natural Science students, and low percentage of counter-stereotypical traits, to be representative of high perceived group homogeneity (Park & Judd, 1990; Boldry, 2007). Therefore, in the PER task, we asked participants to rate each group on four stereotypical and four counter-stereotypical traits, in addition to two neutral traits. We asked participants to rate how many members of each of the two groups they

expect to have the 10 traits. The participants answered on a scale from 0% to 100%, where 0% means that no members in the group possess trait X and 100% means that all members of this group possess trait X.

2.6.3.4 Rating Group Members (RGM) Task

This task consisted of 11 trials (one trial for each presented student). In each trial, the participants were shown a photo of each group member and the group to which the individual belonged. Afterwards, the photo disappeared, and we asked the participants to indicate the group membership of each student. Then the participants were asked to rate how well stereotypic, counter-stereotypic, and stereotype-irrelevant traits described the student whose portrait was shown. The rating was done on a five-point scale ranging from 1 (not at all) to 5 (very much). The used traits were the same as in PER. The question asked was: “Wie sehr beschreiben die folgenden Adjektive die gezeigte Person”.

2.6.3.5 Group Identification Questionnaire

This questionnaire consisted of 16 questions to measure participants’ identification with their social group, adapted from Pickett & Brewer 2001. In this questionnaire, participants were asked to indicate how much they agreed with the shown statements about their group membership on a six-point scale ranging from A (strongly disagree) to F (strongly agree). We used a letter-based scale similar to that described by Pickett and Brewer (2001). Examples of the statements for Natural Science students were, “Ich bin sehr interessiert daran, was andere über Naturwissenschaftsstudent*innen denken.”, and “Wenn ich über Naturwissenschaftsstudent*innen spreche, sage ich üblicherweise “wir” statt “sie”. Four questions out of the 16 were reverse-scored.

2.6.3.6 Mood Scale Questionnaire

The mood scale that participants were asked to complete was a version of (Watson et al., 1988) positive and negative affect scale (PANAS). Twenty-eight items were included in order to measure the valence of participants’ mood (positive or negative) and also the level of arousal produced by the manipulations (high or low). Participants were asked to rate to what extent they feel each of these emotions by using a 1 to 10 scale ranging from (I) do not feel this way at all to (10) feel this way extremely.

2.6.3.7 Manipulation Check Questionnaire

We asked the participants whether they personally knew any of the students whom we presented in the experiment. Additionally, because our intervention involves deception, we asked participants about their suspicions regarding the SAQ feedback or any other aspect of the experiment. The questions included feedback about the participants' own scores or the scores of the ingroup and the outgroup. With this, we checked whether our manipulation was perceived as we intended. We also asked how the participants perceived the study itself, what they thought the purpose of the study was, and whether they perceived anything odd or suspicious in the study. The questionnaire consists of eight questions in total (see Appendix 12.12).

2.6.3.8 Demographics Questionnaire

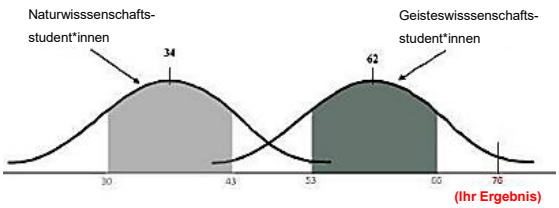
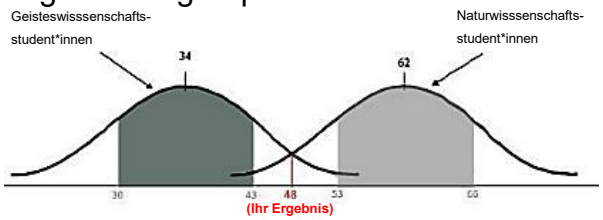
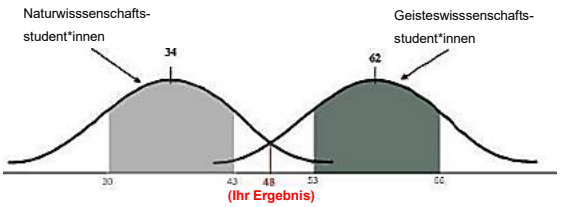
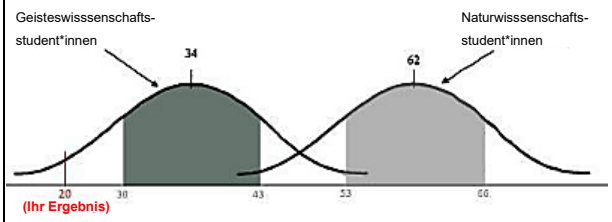
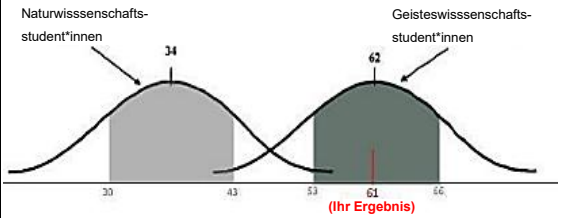
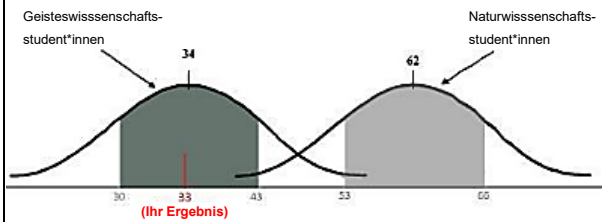
standard demographic questions about participants' gender, age, study field, and highest degree of education

2.6.4 Experimental Conditions

We manipulated the SAQ score results to produce a social identity threat condition or a no-threat condition. However, there were two experimental factors that we had to account for in our paradigm. First, the *Ingroup status,* which is a relative measurement of the status of the ingroup in comparison to the outgroup. The ingroup status could influence the perceived perception of group homogeneity (Ostrom & Sedikides, 1992; Pickett & Brewer, 2001). This was found, for example, in majority vs minority group judgements of perceived group homogeneity. For the SAQ, having a higher score means that individuals have better insight and understanding of themselves. For this reason, the group that had a higher average SAQ score was considered to have a higher status group in comparison to the other group. The second factor was *Within-group status,* which is the status of the participant in comparison to the average status of their ingroup. Thus, within-group status would be high if the participant's SAQ score is higher than their group average, regardless of the status of the group itself. Vice versa, low within-group status would be when the SAQ score of the participant is lower than their group average. Within-group social status could also influence the perception of social identity and social group homogeneity. Having multiple factors, we tried to counterbalance these factors by creating six experimental conditions as follows: three conditions with high ingroup status a) identity threat with high

within-group status, b) identity threat with low within-group status, and c) control condition (no identity threat). The three other conditions followed the same logic, but with having a low ingroup status.

Table 2: shows the experimental conditions in Study 1. The experiment included six conditions. The SAQ feedback for each condition is shown in the corresponding cell.

	High ingroup status	Low ingroup status
Social identity threat	High within-group status:  Participant's score: 76 Ingroup average: 62 Outgroup average: 34	High within-group status:  Participant's score: 48 Ingroup average: 34 Outgroup average: 62
Social identity threat	Low within-group status:  Participant's score: 48 Ingroup average: 62 Outgroup average: 34	Low within-group status:  Participant's score: 20 Ingroup average: 34 Outgroup average: 62
Control	Control within-group status:  Participant's score: 61 Ingroup average: 62 Outgroup average: 34	Control within-group status:  Participant's score: 33 Ingroup average: 34 Outgroup average: 62

2.6.5 Participants

A total of 41 (28 female, 13 male) participants were recruited for the study through social media and academic groups. Requirements for participation were an identification with one of the investigated groups and a signed consent form. Among the participants, 33 belonged to the group AH and 8 to the group NAT. The mean age was 25-26, with a total range from 19 to 39 years. Most of the participants (N = 22) were undergraduate students, while the others were postgraduate students. Five participants had already finished with their master's (4) or doctor's (1) degrees.

2.6.6 Experiment Flow (Procedure)

2.6.6.1 Introduction

The experiment was conducted entirely online in the following manner. After recruitment, participants received a personal ID and URL to access the online study over Qualtrics (Qualtrics, Provo, UT). They were asked to enter a personal ID at the beginning of the experiment. Participants were told that the study investigates the perception of Arts and Humanities Students and Natural Sciences Students. The participants were then told about the study flow.

2.6.6.2 Pre-Intervention Phase

The participants filled out the Mood Scale Questionnaire and the Group Identification Questionnaire to assess the level of ingroup identification before moving on to the SAQ. Afterward, they were shown the Introductory Videos of in- and outgroup members as well as law students. The videos were preceded by labels that indicated to which group the following member belonged. They then proceeded to the first set of homogeneity measures (SIM, PER, and RGM). The participants completed the SIM task followed by the PER Task, once for the ingroup and once for the outgroup. Following that, the participants conducted the RGM task for ingroup, outgroup, and control students.

2.6.6.3 Intervention (SAQ feedback)

After completing the first set of homogeneity measures, the participants were shown their SAQ score as a single number, along with a description of how students who belonged to either the AH or NAT groups performed. In all six experimental conditions, we told the participants that the scores of ingroup students and outgroup students were clearly

different. We employed a between-subjects experimental design, in which the participants' personal scores and the average scores for the ingroup and outgroup were manipulated to produce one of the six conditions for each participant. The scores were shown in numbers and a graphical (histogram) presentation to clearly depict the difference between the participant's score and ingroup and outgroup scores. The experimental conditions are shown in Table 2. Additionally, the participants were instructed to take their time reading the feedback, as the presentation of the feedback would also serve as a break during the study. We expected that this approach would allow the information to settle in, and more time would have passed since answers were given for the first set of measures, enabling participants to have as little memory as possible of their previous answers.

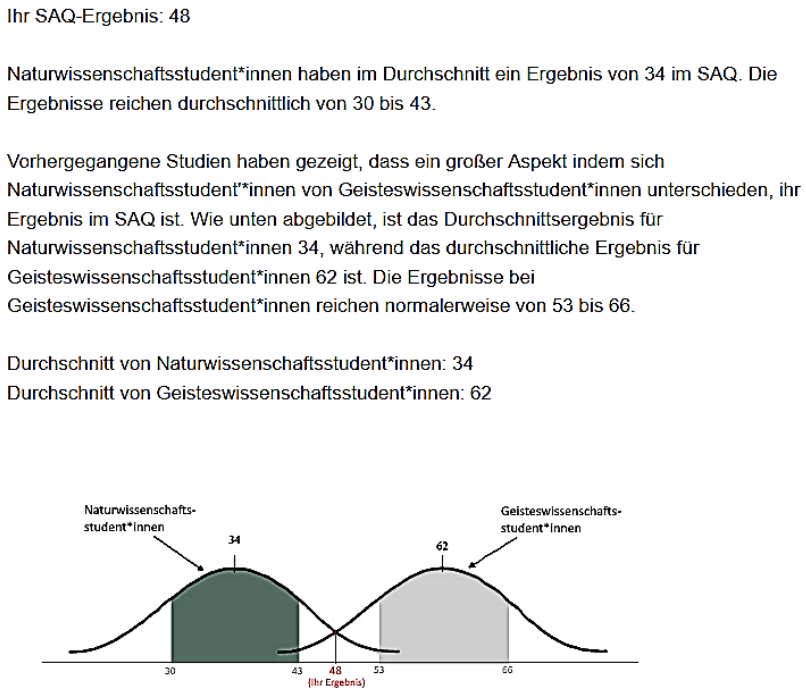


Figure 1: An example of the intervention (SAQ feedback) where NAT group is shown to have a lower status than AH group. The participant is a NAT student, so they had a low group status and a high within-group status. Since the participant's score is not similar to the average score of their ingroup, this condition is considered an "Identity Threat" condition.

2.6.6.4 Post-Intervention Phase

After reading the feedback forms, the participants were shown the introductory videos again and proceeded to the second set of homogeneity measures. The second set of homogeneity measures consisted of the same three questionnaires (SIM, PER, and RGM)

but in a different order, with randomized presentation of the traits. Finally, the participants were asked to complete the Mood Scale Questionnaire again, answer a simple demographic questionnaire, and complete the Manipulation Check Questionnaire. We asked whether any participant knew any of the individuals shown in the videos personally - this was never the case. The study took approximately 70 minutes to complete.

2.6.7 Measures

2.6.7.1 Perceived Group Similarity

For each participant, the perceived group similarity in terms of personality and social skills for the ingroup and outgroup was assessed both before and after the intervention, resulting in a total of 8 variables per participant.

2.6.7.2 Perceived Group Stereotypicality

For each participant, perceived group stereotypicality equals the average percentage of the four stereotypical traits minus the average percentage of the four counter-stereotypical traits. There was one value for the AH group and one for the NAT group.

2.6.7.3 Group Stereotype-Relevant Score

For each participant, we calculated the average percentage of the four stereotypical traits from RGM Task ratings. This resulted in eleven scores, one for each student presented in the introduction videos.

2.6.7.4 Group Stereotype-Irrelevant Score

For each participant, we calculated the average percentage of the two stereotype-irrelevant traits from the RGM Task ratings. This resulted in eleven scores, one for each student presented in the introduction videos.

2.6.7.5 Individual Stereotypicality

We obtained eleven ratings from each participant, one for each student in the introductory videos. It was calculated similarly to Group stereotypicality by subtracting the average score of counter-stereotypical traits from stereotypical traits.

2.6.7.6 Group Identification Level

We calculated the score from the group identification questionnaires from the 16 questions. Then we did a median split based on the identification score, where each

participant was classified either as a “Low Identifier” or a “High Identifier”. The number of participants for the median split was 26 due to the exclusion criteria.

2.6.7.7 Mood Scale Ratings

The 28 adjectives used in the Mood Check Questionnaire were categorized into four dimensions based on the 2D valence-arousal model: positive valence–high arousal, positive valence–low arousal, negative valence–high arousal, and negative valence–low arousal. Each category contained seven traits. The average ratings were then calculated for each category.

2.7 Data Analysis

2.7.1 Analysis Software

The raw data for the present study were acquired using Qualtrics (Qualtrics, Provo, UT) and processed using MATLAB version 9.10.0 (R2021a; The MathWorks Inc., Natick, Massachusetts) to create the dependent variables. The statistical tests were conducted using JASP (JASP Team, 2020, version 0.14.1) and SPSS (IBM Corp., 2020, IBM SPSS Statistics for Windows, Version 27.0, Armonk, NY: IBM Corp.).

2.7.2 Data Preprocessing

Because we were interested in ingroup vs outgroup perception, and our participants were from both Natural Science and Arts and Humanities groups, we had to rearrange the measurements. For participants studying natural sciences, e.g., Biology or Physics, ratings about the NAT group or students were considered ingroup ratings. On the other hand, ratings about the AH group or students were considered outgroup ratings. The same logic was applied to measurements collected from students studying Arts and Humanities topics, e.g., History or Politics. The next step was checking for data normality. Before conducting statistical tests for each hypothesis, the assumptions for each planned test were checked. Therefore, assumptions of normality via the Shapiro-Wilk test and assumptions of equality of variances via Levene’s test were conducted before hypothesis testing.

2.7.3 Mood Check Questionnaire

First, we conducted a bivariate Shapiro-Wilk test to test for the normality of the distributions. The test showed that data of negative valence-positive arousal and negative valence-negative arousal possibly originated from a non-normal distribution (W=0.924,

p=0.009; W=0.936, p=0.024, respectively). So, we chose to perform a non-parametric test (Wilcoxon signed-rank test).

2.7.4 Hypothesis 1

After discovering a discrepancy between our similarity task and Pickett and Brewer 2001, we excluded the ratings of similarity in social skills. This is because, in the original study, they asked the participants to rate the similarity in “Social Life”, not “Social Skills”. Shapiro-Wilk Test showed a significant departure from normality for similarity ratings of the personality, and stereotypicality ratings as assessed in the Percentage Estimate Task (W = 0.880, p < .001; W = 0.904, p = 0.002, respectively). Assumptions were met for the non-parametric Wilcoxon signed-rank test.

2.7.5 Hypothesis 2

Not all of our 41 participants managed to recall the group membership of the individual students. Therefore, we used an exclusion criterion where we excluded participants who failed to identify at least two members (out of 4) in each group. The number of participants included was 23 (16 females), with a mean age of 25.8 years. Because we were interested in the correlation between PER results and RGM results, we conducted Shapiro-Wilk tests for bivariate normality. The test revealed that the distributions of ingroup ratings of PER and RGM significantly depart from normality (W = 0.828, p = .001). On the other hand, the distributions for outgroup ratings were derived from a normal distribution (W = 0.964, p = 0.553). Since some of the distributions were not normally distributed, we opted to use Kendall’s Tau test to measure the correlation.

2.7.6 Hypothesis 3

To test whether inducing social identity threat influences perceived outgroup homogeneity. As we tried to replicate the experiment done by Pickett & Brewer (2001) as closely as possible, we only included Arts and Humanities students We also excluded participants who expressed suspicion in SAQ scores feedback and participants who did not recall their SAQ score within a 5-point range and participants who did not perceive the two groups to be different to each other in SAQ scores. In total, we had 26 Arts and Humanities students included, all were bachelor’s degree students or recent graduates. We conducted the data analysis in two methods. The first method was done by replicating what Pickett & Brewer did. However, because we also have pre-intervention measurements, we thought that we

might have more comprehensive data to adjust for inter-individual differences in pre-intervention ratings.

2.7.6.1 Hypothesis 3.1

Outlier Detection and Data Exclusion: We examined the data for outliers by eye to determine whether data from some participants could have biased our parametric statistical test. To detect outliers, we created boxplots for our dependent variable values. A total of 8 outliers in the similarity ratings and 11 outliers in the percentage ratings were visible in the boxplot diagrams. We looked for any patterns in answers that could indicate that exceptionally low or high ratings were not valid. For example, if all ratings had a value of 3, independent of ingroup or outgroup ratings. Individual assessment of the participant's ratings showed no such patterns in the ratings. Additionally, some of these participants indicated in the manipulation check questionnaire their interest in the study. The outliers were thus included in the analysis. To replicate Pickett & Brewer's 2001 analysis, we used only post-intervention ratings from PER, SIM, and RGM tasks. Pickett & Brewer (2001) performed analyses of variance on their data and used the type of intervention (threat vs. no-threat), ingroup status (high vs. low), and identification level (high vs. low) as fixed factors and separately, the similarity and percentage ratings as repeated measures dependent variables. Due to our low number of participants, we were unable to perform the same analysis and therefore had to exclude the ingroup status as a fixed factor.

Assumption Testing for Two-Way ANOVA: First, we tested whether the assumptions of two-way ANOVA were met. PER and SIM observations were independent, the independent variables (2 x 2 ANOVA factors) were categorical, and the dependent variables (PER and SIM ratings) were continuous. However, we needed to test the two assumptions of the two-way ANOVA test. First, the Normal distribution of data, using the Shapiro-Wilk test, and second, the homogeneity of variance for each combination of tested groups using Levene's test. Shapiro-Wilk Tests were significant for ingroup similarity ratings on the personality dimension from participants of the no-threat condition ($W = 0.775$, $p = .007$) and for outgroup similarity ratings on the personality dimension from participants of the threat condition ($W = 0.858$, $p = 0.018$), rejecting the normality assumption. All other combinations of the groups were statistically not significant in the Shapiro-Wilk Tests, i.e., the samples most likely derived from a normal distribution. We used Levene's test to test for homogeneity of variance. Levene's test was conducted on

each combination of the groups of the independent variables. A comparison of the threat vs. no threat condition revealed significant results in outgroup ratings for the Similarity Task and the Percentage Estimate Task on stereotype-relevant ratings ($F = 7.047$, $p = 0.014$; $F = 7.271$, $p = 0.013$, respectively). The variances in these ratings under threat vs. no threat conditions were thus statistically not homogeneous.

Statistical Test Selection: Because similarity ratings of personality violated the two assumptions of the two-way ANOVA, we performed a non-parametric two-way factorial design based on the Kruskal-Wallis with a Schreier-Ray-Hare extension for the Similarity Task. Data from PER and RGM were analyzed with a two-way repeated-measures ANOVA.

Analysis: We used a 2 (condition: threat vs. no-threat) x 2 (group identification: low vs. high) multifactorial design to test for the effect of identity threat and group identification on Similarity Ratings in the personality dimension. This was done for the ingroup and the outgroup separately. For Percentage Estimate ratings, we used a 2 (condition: threat vs. no-threat) x 2 (group identification: low vs. high) x 2 (stereotype relevance: relevant vs irrelevant) multifactorial analysis of variance with repeated measures on the factor stereotype relevance since these ratings were derived from the same participant. This was done for ingroup and outgroup data separately.

2.7.6.2 Hypothesis 3.2

For our own analysis, we wanted to adjust for baseline ratings in all tasks (pre-intervention ratings). So, our tests had to be pairwise comparison tests. Our dependent variables were similarity ratings in personality and group stereotypicality scores as calculated from the PER task before and after the intervention. The ratings were separated into ingroup and outgroup ratings and threat vs. no threat conditions, resulting in four data distributions for each measure.

Assumption Checks: *First,* we tested for the normality of each data distribution. The Shapiro-Wilk test was only significant for ingroup and outgroup ratings of similarity in personality in the no-threat group (Ingroup: $W = 0.781$, $p = 0.008$; Outgroup: $W = 0.774$, $p = 0.007$).

Statistical Test Selection: We performed a single-factor ANOVA (threat vs. no threat) followed by a non-parametric post-hoc test (Kruskal-Wallis test) for SIM task ratings. For group stereotypicality ratings, we conducted an ANCOVA with identity threat as a fixed

factor and group identification score as a covariate. Then the ANCOVA was followed by a post-hoc test for PER stereotypicality scores.

Change in Ratings: For the two groups we had (Identity threat vs. no-threat), we subtracted pre-intervention ratings from post-intervention ratings for both SIM and PER tasks. Our new measures represent the change in ratings from before to after the intervention. We had two measures: a. change in similarity ratings of personality, and b. change in group stereotypicality.

Assumption Checks for Differences: Shapiro-Wilk Tests showed a significant deviation from normality for the change in SIM task ratings for ingroup and in the no-threat condition (Ingroup: $W = 0.781$, $p = 0.008$; Outgroup: $W = 0.774$, $p = 0.007$). Similarly, stereotype-relevant ratings of the outgroup from the PER task in both conditions were significant in the Shapiro-Wilk test (Threat: $W = 0.850$, $p = 0.013$; No-Threat: $W = 0.761$, $p = 0.005$). Therefore, we conducted non-parametric Mann-Whitney U tests on all comparisons to better compare the resulting data.

2.8 Results:

2.8.1 Manipulation Check

In the Manipulation check questionnaire, participants were asked to recall their own SAQ score and the average SAQ score of the ingroup and the outgroup. 90.24 % of the participants recalled their own score within 5 points of error. While only 82.93 % and 73.17 % of the participants recalled the average ingroup and outgroup scores, respectively, within 5 points of error, most of the participants (95,12 %) also recalled their relative score in comparison to both ingroup and outgroup averages. In addition, 36 participants (87,8 %) correctly recalled the relative group status of the ingroup relative to the outgroup. In general, the results suggest that the participants paid attention to the information given on the SAQ feedback form. Participants who did not recall the group's average SAQ scores as different from each other (N=5) were excluded from the analysis of the manipulation effect. Generally, outgroup averages were recalled by fewer participants than ingroup averages or their own scores.

2.8.2 Mood Check Questionnaire

Mean mood rating post-intervention ranged from 2.129 (SD = 1.756) for the negative valence–high arousal dimension to 5.174 (SD = 1.036) for the positive valence–low

arousal dimension. Wilcoxon signed-rank tests showed a significant increase in positive arousal-related emotions: Negative valence-positive arousal emotions and positive valence-positive arousal emotions showed a significant increase after the intervention (respectively: W=498, p=0.01; W=485, p=0.017). The participants generally reported higher ratings for positive valence adjectives than for negative valence adjectives.

2.8.3 Group Identification

A median split of the average score of each participant in the group identification questionnaire was performed to classify participants into high or low identifiers. The median across all participants (N = 41) on the 6-point scale was 4.0, with a mean value of 3.9 (SD = 0.732). After the median split, we had 21 low identifiers and 20 high identifiers.

2.8.4 Hypothesis 1

We tested this hypothesis only using pre-intervention data so that participants could express their general perception of the ingroup and outgroup before we manipulated their sense of social identity through the SAQ test feedback. Wilcoxon-signed rank showed that similarity ratings in personality were significantly higher for the outgroup than for the ingroup (W= 69.0, p= 0.009). Then again, the participants rated the group stereotypicality of the outgroup to be higher than the group stereotypicality of the ingroup (W=150, p=0.002), see Figure 3. Overall, the results showed significantly higher perceived similarity in personality of outgroups compared with ingroups and significantly higher perceived outgroup stereotypicality in comparison with ingroups. Therefore, hypothesis 1 was confirmed.

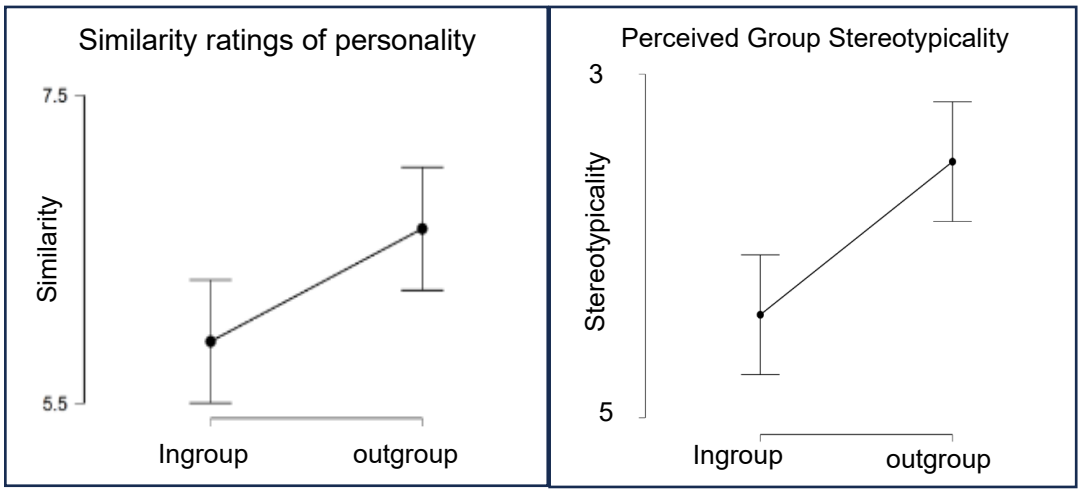


Figure 3: Post-intervention ratings of similarity in personality and group stereotypicality. The left subplot shows the average similarity ratings of personality for ingroups and outgroups, with the error bar indicating standard error. The right subplot shows calculated group stereotypicality for both ingroup and outgroup. The Perceived Group Stereotypicality was calculated as the difference between the average ratings on stereotypical traits minus the average rating on counter-stereotypical traits.

2.8.5 Hypothesis 2

For the correlation analysis, only participants who recalled at least two individuals from each group correctly were considered (N = 23). Correlation testing suggests that the Group stereotypicality of the ingroup (calculated using the PER task) positively correlated with the average stereotypicality of the ingroup individuals (calculated using the RGM task). Significant correlation (T_b= 0.486, $p < 0.001$). However, group stereotypicality of the outgroup did not significantly correlate with the average stereotypicality ratings of the outgroup individuals (T_b= -0.148, $p = 0.836$). Therefore, our second hypothesis was only confirmed for ingroups, where individual-level stereotypicality positively correlated with group-level stereotypicality.

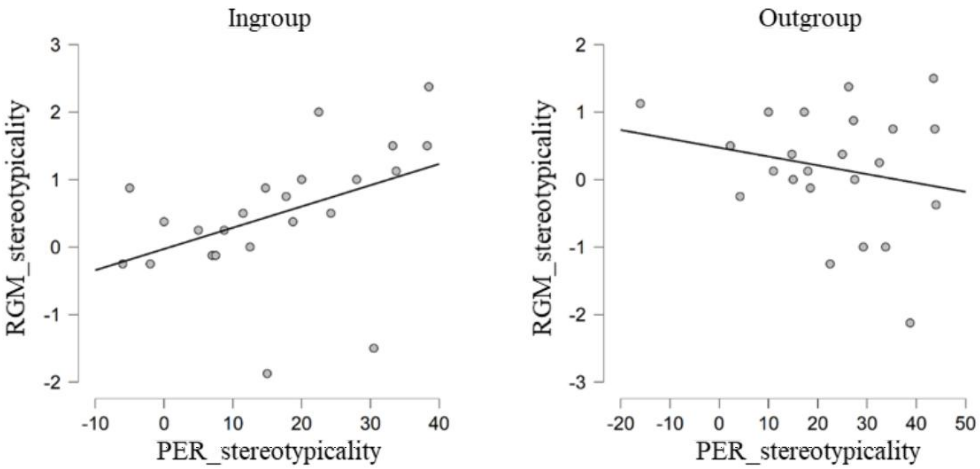


Figure 4: Correlation between the average Individual-based Stereotypicality calculated from the RGM task and Group-based Stereotypicality calculated from the PER task. Each dot indicates the calculated rating from one participant (n=21).

2.8.6 Hypothesis 3

2.8.6.1 Hypothesis 3.1

In hypothesis 3, we tested whether social identity threat induced perceived out-group homogeneity. We used post-interventional ratings in the first iteration of hypothesis 3 (3.1) to be as close to the analysis done by Pickett & Brewer (2001).

2.8.6.1.1 Similarity Ratings of Personality:

Table 3: Descriptive statistics of Similarity Ratings of Personality (SIM) divided into four relevant groups (2 threat conditions vs. 2 group identification statuses).

Condition	Number	Similarity in personality Mean (Standard Deviation)	
		Ingroup	Outgroup
Threat / High identifier	8	6.9 (2.6)	7 (1.9)
Threat / Low identifier	8	5.6 (1.8)	5.5 (2.6)
No-threat /High identifier	5	7.8 (0.8)	7.8 (0.8)
No-threat /Low identifier	5	5.4 (2.7)	7 (0.7)

For ingroup ratings, the analysis included two factors: Identity Threat (no threat vs. threat) and Group Identification (high identifier vs. low identifier). Results revealed a significant main effect of group identification, $F_{(1,22)} = 4.305$, p=.050, indicating that high identifiers (mean = 7.34) reported higher ingroup similarity compared to low identifiers (mean = 5.50).

The main effect of identity threat was not significant, $F_{(1,22)}$ =0.158, p=.695, suggesting no difference in perceived similarity between the no-threat and threat conditions. Additionally, the interaction between identity threat and group identification was not significant, $F_{(1,22)}$ =0.427, p=.520, indicating that the relationship between identity threat and in-group similarity did not depend on the level of group identification.

For outgroup ratings, results of the two-way ANOVA indicated that the main effect of Identity Threat was not significant, ($F_{(1,22)}$=2.3, p=0.14), suggesting no difference in out-group similarity between the no-threat and threat conditions. Similarly, the main effect of Group Identification was not significant, $F_{(1,22)}$ = 2.3, p=0.14. Additionally, the interaction between Identity Threat and Group Identification was not significant, $F_{(1,22)}$=0.216, p=0.65.

The Schreier-Ray-Hare extension on the Kruskal-Wallis test was performed to assess the effect of identity threat and group identification on the similarity ratings for the ingroup and outgroup. The results were significant for the effect of group identification on ingroup similarity ratings only ($H_{(1)}$ = 4.65, p = 0.03)

Percentage Estimate Task (PER) Ratings: A repeated measures ANOVA was conducted to examine the effects of Stereotype Relevance (relevant vs. irrelevant), Identity Threat (no threat vs. threat), and Group Identification (high identifier vs. low identifier) on in-group similarity ratings. The analysis included within-subjects effects for stereotype relevance and its interactions and between-subjects effects for identity threat, group identification, and their interaction.

Within-Subjects Effects: The analysis revealed a significant main effect of Stereotype Relevance, $F_{(1,22)}$=92.851, p<.001. However, no other significant main effects were found. In addition, all of the interactions between the factors were not significant.

Figure 5: Descriptive statistics comparing PER results of stereotype-relevant vs. stereotype-irrelevant traits. The figure shows that participants reported higher in-group similarity when the stereotype was relevant (e.g., M=73.70, SD=7.087 (for high identifiers in the no-threat condition) compared to when the stereotype was irrelevant (e.g., M=33.30, SD=15.283) for high identifiers in the no-threat condition). The trend of higher ratings under relevant stereotypes was consistent across threat and group identification levels.

These results indicate that Stereotype Relevance significantly affects ingroup similarity ratings, but its interactions with Identity Threat and Group Identification do not reach statistical significance.

In conclusion, using a similar hypothesis testing to the original study by Pickett & Brewer 2001, we did not find a significant effect of Identity threat in both SIM and PER tasks for ingroup or outgroup ratings. Therefore, we could not replicate the finding of Picket & Brewer 2001 that participants who were exposed to social identity threat perceived the outgroup to be more homogenous than control participants.

2.8.6.2 Hypothesis 3.2

With our study design, we adopted a within-subject design to measure the change in similarity ratings of personality and group stereotypicality ratings from before to after the intervention. Therefore, in the second iteration of hypothesis 3 (3.2), we tested whether participants who received a social identity threat intervention showed a bigger change in ratings of similarity and stereotypicality from before to after the intervention than participants in the control condition (no threat). In order to measure that, we subtracted pre-intervention measures from post-intervention measures and used the differences/changes in measures as our dependent variables.

Change in Similarity Ratings of Personality (SIM Task): The change in scores was analyzed using ANOVA with Identity Threat as a factor. The ANOVA revealed that the main effect of Identity Threat approached significance ($F_{(1, 23)} = 3.65$, $p = 0.068$), suggesting a

trend toward a difference in change scores between the no-threat and threat conditions. The Kruskal-Wallis test showed a marginal effect but not a statistically significant effect on Identity Threat ($H_{(1)}$ = 3.566, p = 0.059).

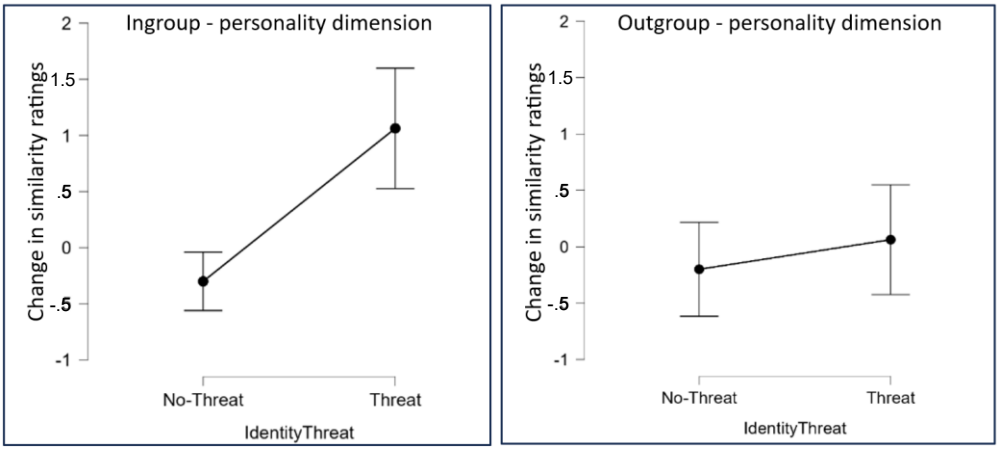


Figure 6: The change in similarity ratings of personality. The left subplot exhibited a small decrease in ingroup similarity ratings (M = -0.3, SD = 0.8, N = 10), while participants in the Threat condition showed an increase (M = 1.1, SD = 2.1, N = 16) with greater variability in the threat condition. For outgroup ratings of similarity, the ANOVA revealed no significant main effect of Identity Threat ($F_{(1, 23)}$ = 0.140, p = 0.711), indicating that changes in outgroup similarity ratings were not significantly different between the no-threat and threat conditions. Kruskal-Wallis Test confirmed no significant effect of Identity Threat ($H_{(1)}$ = 0.012, p = 0.913).

Change in Perceived Group Stereotypicality (PER Task): For ingroup stereotypicality ratings, the ANCOVA revealed no significant main effect of Identity Threat, $F_{(1,23)}$=0.04, p=0.84, or group identification score ($F_{(1,23)}$=0.27, p=0.6) indicating that changes in ingroup stereotypicality ratings were not significantly different between the no-threat and threat conditions. A post-hoc t-test revealed no significant difference between the no-threat and threat conditions (t=0.2, p_{Holm}=0.85). For outgroup stereotypicality ratings, the ANCOVA revealed no significant main effect of Identity Threat, $F_{(1,23)}$=0.63,p=0.43, or group identification score ($F_{(1,23)}$=2.27, p=0.14) indicating that changes in outgroup stereotypicality ratings were not significantly different between the no-threat and threat conditions and didn't depend on the level of group identification. Post-hoc t-test revealed no significant difference between the no-threat and threat conditions (t= -0.8, p_{Holm} = 0.43).

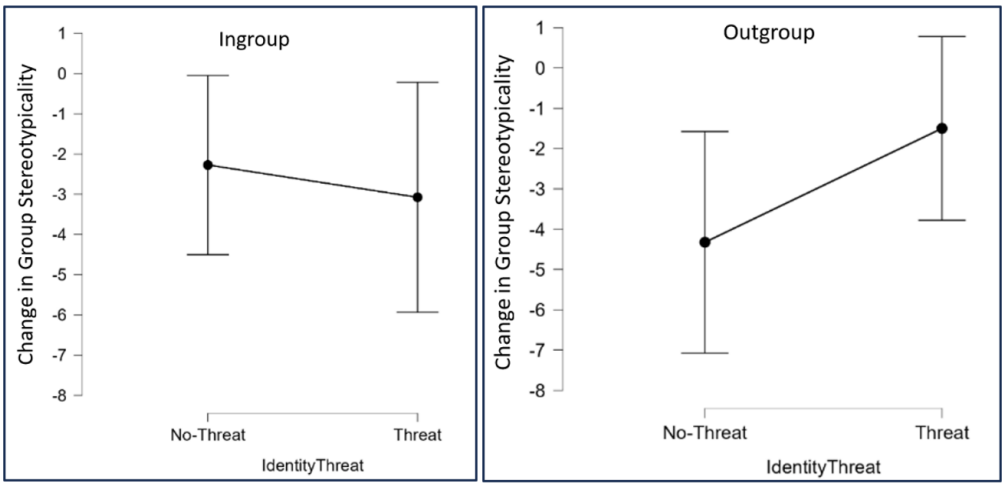


Figure 7: The change in Perceived Group Stereotypicality for ingroup (left subplot) and outgroup (right subplot) in the No-threat and Threat conditions. There was a small but highly variable decrease in perceived stereotypicality ratings of the ingroup in the no-threat condition (M = −2.3%, SD = 7%, N = 10), and in the threat condition (M = -3%, SD = 11%, N = 16). Similarly, a small and highly variable decrease was found in perceived outgroup stereotypicality in the no-threat condition (M = -4.3%, SD = 8.6%, N = 10) and threat condition (M = −1.5%, SD = 9%, N = 16).

In conclusion, using a different analysis method where we subtracted pre-intervention ratings of ingroup and outgroup, we found an effect of identity threat on perceived ingroup and outgroup homogeneity. However, this effect was limited only to the similarity ratings of personality for the ingroup. This showed that compared with the control condition, participants under social identity threat changed their views of their own group and perceived its members to be more similar in personality.

Overall, we couldn't confirm hypothesis 3, which postulated that social identity threat should increase perceived outgroup homogeneity, which is a prediction of ODT theory and was previously reported by Pickett & Brewer 2001.

2.9 Discussion

This study aimed to replicate and extend prior findings on perceived group homogeneity, with a specific focus on how social identity threat influences perceived similarity and stereotypicality within and between social groups. Using a modified version of the paradigms introduced by Pickett and Brewer (2001), we aimed to evaluate the validity of group-based versus individual-based measures of group homogeneity and to assess their responsiveness to identity threat interventions.

2.9.1 Summary of Main Findings

The study's hypotheses were tested through a combination of group-level (PER and SIM tasks) and individual-level (RGM task) measures.

Hypothesis 1, which predicted higher similarity and stereotypicality ratings for the outgroup compared to the ingroup, was partially supported. Participants rated the outgroup as significantly more similar in terms of personality and more stereotypical than the ingroup. However, this effect did not extend to social skills, likely due to methodological divergence from the original "social life" framing used in Pickett and Brewer's study.

Hypothesis 2 posited a positive correlation between group-level and individual-level stereotypicality ratings. Results supported this for the ingroup, but not for the outgroup, suggesting that aggregated group judgments may not reliably reflect perceptions of individual outgroup members. These findings challenge assumptions that group-level measures are sufficient proxies for individual-based cognition and align with previous literature (Ostrom & Sedikides, 1992; Park & Judd, 1990) that highlights the variability of results based on the chosen measure.

Hypothesis 3, which explored the effect of social identity threat on perceived group homogeneity, yielded mixed results. Although descriptive trends hinted at increases in perceived homogeneity under threat, statistical tests (ANOVA and nonparametric alternatives) revealed no significant differences between the threat and control groups. Neither ingroup nor outgroup stereotypicality ratings changed meaningfully as a result of the social identity threat, and the level of group identification did not moderate these effects.

2.9.2 Interpretation and Theoretical Implications

These findings partially replicate core effects in the social perception literature, particularly concerning outgroup homogeneity and the non-equivalence of group-based and individual-based judgments. The observed lack of significant effects for identity threat may reflect limitations in the strength or believability of the manipulation. In addition, because the experiment was conducted online, the participants might not have been emotionally involved in the experimental paradigm sufficiently to the degree of feeling identity threat.

Importantly, our results highlight the role of measurement type in influencing the observed effects. For instance, group-based measures appeared more sensitive to general stereotypes, while individual-based measures were less consistent, especially for the

outgroup. This difference should drive researchers to make a distinction between individual-based and group-based measures.

Moreover, our inability to replicate all findings from Pickett and Brewer (2001) underscores the contextual and cultural variability of social psychological effects. We conducted our experiment 21 years after the original experiment. In addition, our sample—limited to Arts and Humanities students from Germany—may differ from prior populations in meaningful ways (e.g., the salience of academic identity, cultural norms, and previous beliefs about the NAT and AH study fields around intergroup judgment), which may have affected the strength of the identity threat.

As mentioned before, previous social psychology work relied on tasks asking participants to rate the group as a whole. However, asking about specific individuals within the group might deviate ratings toward these individuals, making them more salient or, worse, leading to complete individualization where participants rate individuals independently of the group. On the other hand, judging the group as a whole, particularly in natural groups, introduces unmeasurable confounders. It becomes difficult to determine what participants are thinking of when rating the group as a whole.

2.9.3 Methodological Considerations and Limitations

Several methodological factors warrant consideration. First, our pre-intervention ratings helped control for individual baseline differences, yet this design diverged from previous work and may have influenced the overall comparability of results. Second, although we preserved much of the original study design, some adaptations (e.g., different phrasing in similarity tasks, and sample characteristics) may have diluted the effect of the manipulation. Third, the small sample size, particularly after applying strict inclusion criteria (e.g., manipulation check pass, accurate recall), reduced the statistical power.

2.9.4 Future Research Directions

Future studies could build on these findings by (1) increasing the emotional salience of the identity threat manipulation, (2) incorporating longitudinal or repeated exposure designs to assess the persistence of homogeneity perceptions over time, and (3) expanding the sample to include more diverse participant pools (e.g., cross-cultural or mixed-discipline groups). Moreover, neuroimaging approaches, such as those planned for subsequent studies in this thesis, offer promising avenues to explore the neural correlates

of perceived group homogeneity and stereotype processing. This could help uncover whether individual-level neural responses reflect the divergence observed between group- and individual-based ratings.

3 Study 2: Behavioral Study Inducing Social Categorization

3.1 Introduction

3.1.1 Group Perception from a Similarity Perspective

Previous literature on outgroup homogeneity and current neuroimaging research on social perception converge on an important aspect: “similarity in representations.” Since Tajfel and Turner’s introduction of social categorization and social identity theories, the concept of homogeneity has been used to describe how similar group members are perceived to be. More generally, identifying with an ingroup depends on perceived similarities between oneself and the group members. From this perspective, social categorization can be seen as a way of clustering similar stimuli together and distancing dissimilar ones. Perceived group homogeneity, therefore, becomes a specific case where social group members are perceived as highly similar along a dimension, such as personality traits.

3.1.2 Behavioral Pairwise-Similarity Task

Building on Pickett and Brewer’s (2001) dimensions of personality, social life, and physical appearance, we adopted a pairwise similarity task. This task, which was used by Tsantani et al., (2021), involved participants rating individuals relative to one another across specific dimensions. The Pairwise-similarity task allowed us to quantify in more detail how similar group members were perceived to be relative to each other.

3.1.3 Representing Persons as Stimuli

However, since the concept of a “person” is so general, it would be difficult for the participants to think of every aspect of an individual when judging them. To address this, we developed a neuroimaging task focused on the group membership of the individuals who were being judged. We ensured participants received limited, group-related information about individuals to avoid making certain members more individualized. Representing a “person” as a stimulus posed a challenge, as we sought to avoid overemphasizing specific traits (Freeman & Ambady, 2011).

3.1.4 Faces as a Representation of Persons

For our task, we chose to show participants the faces of the students as they offer the most direct representation of individuals (Yankouskaya et al., 2014) Still, confounders such as gaze, facial expression, and screen position must be controlled (Jack & Schyns, 2015). Additionally, viewing a face does not guarantee that participants think about the "person behind the face." Therefore, we incorporated a task emphasizing group membership for each face to ensure participants were cognitively engaged with this aspect (Castello et al., 2021).

3.1.5 Networks Rather Than Regions

Recent advancements in understanding brain functioning suggest there cannot be a dedicated "brain region for each stimulus category" since multiple regions are typically involved in experimental tasks (Bressler & Menon, 2010; Carrington & Bailey, 2009; Davis et al., 2014; Haxby et al., 2001) For example, studies have identified brain networks like the default mode network, where certain regions function together during specific tasks (Bahrami et al., 2023; Tang et al., 2023; Yao et al., 2015). Neuroimaging studies now focus on neural activation patterns in response to different stimuli rather than on distinct brain regions, which is particularly relevant when dealing with stimuli of the same nature (Popal et al., 2020; Weaverdyck et al., 2020).

3.1.6 Activation Patterns of the Faces and Associated Information

Moreover, presenting faces activates regions responsible for human face processing and general visual processing (Baseler et al., 2014). Literature indicates we should not expect different brain regions to activate for different faces, but rather distinct activity patterns within the same regions (Collins et al., 2016; Collins & Olson, 2014; Kriegeskorte et al., 2007) This aligns with our expectation of consistent neural activation but distinctive patterns for each face, reflecting unique social and emotional associations. For instance, information about group membership can influence how faces are processed in the brain (Van Bavel et al., 2008, 2011)

3.1.7 Difficulties in Measurement of Perceived Group Homogeneity.

From this point on, we had to decide whether we should focus on our experimental paradigms on ingroup vs outgroup perception, or on social group perception in general. We had many reasons to choose to focus on social group perception in general. First of

all, out-group homogeneity seemed to be highly variable between measures and contexts. Several factors were shown to influence perceived group homogeneity, for example, the group status of the participant and the group under judgment (Boldry et al., 2007; Karasawa et al., 2004; Lorenzi-Cioldi, 1993). Personal motivations of the participant, e.g., according to optimal distinctiveness theory, the need for assimilation or differentiation, and these personal motives are also influenced by within-group status (Pickett & Brewer, 2001).

3.1.8 New Study Aim: Focus on the Task

Since there was no previous measure of social group perception using individual-level questions, we decided that it's more important to establish a neural measure that correlates well with our new individual-based behavioral measure (pairwise similarity) and could map the perceived similarities between individuals of a social group whether the group was perceived to be homogenous or not. By this means, we could search for brain regions that show a similarity map in activation patterns that correlate with the behavioral task. In other words, we wanted to develop a behavioral measure of perceived similarity between individuals that could be used to measure all kinds of social groups over any dimension. This brain network, if consistent across multiple dimensions of group-relevant individual characteristics (Whether stereotypical traits or counter-stereotypical traits), e.g., personality, social life, academic life, etc., for Study Field-based groups, could shed light on which neural networks are responsible for social categorization, stereotyping, and comparing persons.

3.2 Study Design and Experimental Procedure

In this study design, differences in activity patterns induced by two faces remain stable unless influenced by new information, such as group membership. After viewing two faces a second time with new group-related information, changes in activity patterns could be attributed to cognitive or emotional associations, such as threat or dislike toward specific groups.

3.2.1 Stimuli

3.2.1.1 Introductory Videos

We used eight videos out of the eleven we recorded. For the sake of simplification, we only included videos of Natural Science students (four videos) and Arts and Humanities

students (four videos), but not the videos of law students, because law students were not reported to belong to either NAT or AH groups in the pretesting we conducted earlier. This is to reduce the number of new persons whom participants need to learn about from eleven to eight. This was also to simplify the comparisons that the participants needed to do, so they only needed to indicate the similarity between persons from two groups rather than three.

3.2.1.2 Color Portraits

The same color portraits that we used in Study 1. The portraits were used in the Group Membership Recall Task (explained in detail in the “Tasks” section).

3.2.1.3 Cropped Greyscale Faces

Eight screenshots were extracted from each student video, then we converted them to grayscale images, and only the faces and hair of individual students were kept in the images and everything else was replaced with a black background. The images were then grayscaled and matched in average luminance and contrast using the SHINE toolbox (Willenbockel et al., 2010) on Matlab 2022a the same procedure as was done by (Castello et al. 2021). Then the face images were rescaled across the diagonal direction so that all images had a similar squared diagonal length in pixels. We included different facial expressions in the images. So, out of the eight images, we made sure that there was at least one image showing the student looking at the camera directly, another one showing the student looking away from the camera, and another image showing the student talking.

3.2.2 Tasks

3.2.2.1 Pairwise Similarity Task

The Pairwise similarity task (Tsantani et al., 2021) consisted of two runs, each one comprised of 56 trials. In each trial, we showed two faces of the eight subjects side-by-side and asked the participants to indicate the level of similarity between the two faces on one or more dimensions. In the first run, the similarity dimension was “physical facial features” and in the second, we asked about two dimensions, namely, “personality” and “social life”. Because we had eight persons, there were 28 possible combinations of any two persons; one person’s face is shown to the right and one to the left. And because we wanted to remove any effect of the face position, we repeated the 28 combinations of persons with a switched position. The stimuli used were the cropped greyscale faces. For

each person viewed, we chose one face per trial randomly from a pool of four cropped faces to represent the person. As mentioned before, the faces were rescaled to have 1000 pixels along the diagonal axis without changing the aspect ratio of the images.

3.2.2.2 Meet the Students Task

In this task, participants viewed eight videos out of the 11 introductory videos we used in Study 1. We only included videos of the four NAT and four AH students. In these videos, the students introduced themselves. Each student mentioned their study field and the reasons why they chose this field.

3.2.2.3 Group Membership Recall Task

This was an essential task in which we tested participants' knowledge of the group membership of the eight students. In this task, the participants first watched a "Memory board" which showed color portraits of members of each group (four in each group). The participants were told to take the time they needed to memorize the group membership of each person. Next, the participants watched the color portrait of each one of the eight students one after the other in a random order and they were required to indicate the group membership of the person they see by pressing right or left keyboard arrows. After answering the test for the eight students, the participants were informed whether they had made mistakes or not. If they answered correctly then they had to do a second round of the test correctly in order to finish the task. However, if the participant had made at least one mistake in the first or the second round of the test, they were required to start the task from the beginning by watching the "Memory board" again and they were asked to finish two rounds of the recall test in order to finish the task. There was no limit on the number of times the participants could repeat the task. However, no participant passed the task except after answering the recall test correctly two times in a row.

3.2.2.4 Group Identification Questionnaire

We used the exactly the same questionnaire as in Study 1.

3.2.2.5 Group Similarity Task (SIM)

In Study 2 and Study 3 we used the same task as in Pickett and Brewer (2001). However, we corrected the mistake made in Study 1, so instead of asking participants to rate the similarity of the students in "Social skills," we asked them to rate their similarity in "Social life." So, In Studies 2 and 3, we asked participants to rate the similarity between group

members of Natural Science Students and Arts and Humanities students in “personality” and “social life”. The task was only performed after the intervention.

3.2.2.6 Percentage Estimate Task (PER)

The participants performed the same task as in Study 1, but only after the intervention.

3.2.3 Experiment Flow

The experiment was done on a Windows 10 laptop using custom Python code in PsychoPy version 2022.1.1 (Open Science Tools, 2022). Participants finished the experiment in around 40 to 60 minutes, according to their pace. For scale-based tasks, the scale marker was always shown first in the center (at number four for a 7-point Likert scale (which was used in SIM and pairwise similarity task) and between “D” and “E” for Group Identification Task) and the scale marker was presented in grey as “inactive”. In order to make a choice, participants were instructed that they must move the scale marker to the right or left first to make it active (the marker turned red). So, even if the participant wanted to choose the rating “4” in the pairwise similarity task, they had to move the marker right or left first to make it active, then choose the number “4”. This restriction was imposed to prevent participants from exploiting the task by just pressing enter to pass to the next slide without thinking enough about the task.

3.2.3.1 Pre-Intervention Phase

Participants first completed the Group identification Questionnaire, then they performed the pairwise similarity task, rating the similarity in “physical facial features” between each of the eight students in 56 slides. After that, participants conducted the pairwise similarity task again rating the similarity between students over “personality” and “social life”.

3.2.3.2 Intervention

The participants conducted the Meet the Students Task and then completed the Group Membership Recall Task. As mentioned earlier, participants had to repeat the recall task until they correctly recalled the group membership of the eight students twice consecutively.

3.2.3.3 Post-Intervention Phase

Directly after the intervention, participants repeated the pairwise similarity task exactly as they had done before. Because the faces for each person were chosen randomly from a pool of faces. We couldn’t control that the post-intervention images were exactly the same

as the pre-intervention images. After that, participants performed the similarity task (SIM) in personality and social life. Then they performed the Percentage Estimate Task (PER) as in Study 1 where they rated both groups on four stereotypical traits, four counter-stereotypical traits, and two stereotype-irrelevant traits.

3.2.4 Measures

Post-intervention similarity values were computed for personality and social skills dimensions, separately for ingroup and outgroup faces, following the same method as in Study 1. Additionally, **post-intervention group stereotypicality** was calculated, yielding one value per group, also using the approach from Study 1. **Perceived within-group similarity (WG)** was defined as the average similarity rating between each pair of faces belonging to the same group. In contrast, **perceived between-group similarity (BG)** was defined as the average similarity rating between pairs of faces from different groups.

3.3 Hypothesis 1

We hypothesized that revealing the group membership of the eight students (the intervention that we hoped would induce social categorization) would lead to an increase in *perceived within-group similarity* and a decrease in *perceived between-group similarity*. This should be manifested as significantly higher within-group similarity ratings than between-group similarity ratings in personality and social life dimensions, but not in physical facial features.

3.4 Pilot Experiment

3.4.1 Participants

We conducted a pilot experiment online using Zoom software (Zoom Video Communications, 2022) in which we shared the screen of the experiment PC with the participant and they were able to control the experiment PC and give their answers. We recruited 16 participants in May and June 2022 for the pilot experiments. Unfortunately, because of bugs in our code, we could not record the pre-intervention pairwise similarity ratings from all the subjects. In addition, we excluded six participants because of missing answers. So effectively, we had ten participants: six Arts and Humanities students and four Natural Science students. The participants were German bachelor students from all over Germany.

3.4.2 Data Analysis

From the pairwise similarity task for each participant, we obtained 16 between-group similarity ratings and 12 within-group similarity ratings (eight ratings of within-group similarity for ingroup and outgroup). So, from each pairwise similarity task for 10 participants, we had 160 between-group similarity values and 120 within-group similarity values. Descriptive statistics of the data showed high skewness of the data distribution, especially in similarity ratings of “physical facial features”. So, we conducted the Shapiro-Wilk test to assess data normality for the three post-intervention pairwise similarity tasks we conducted.

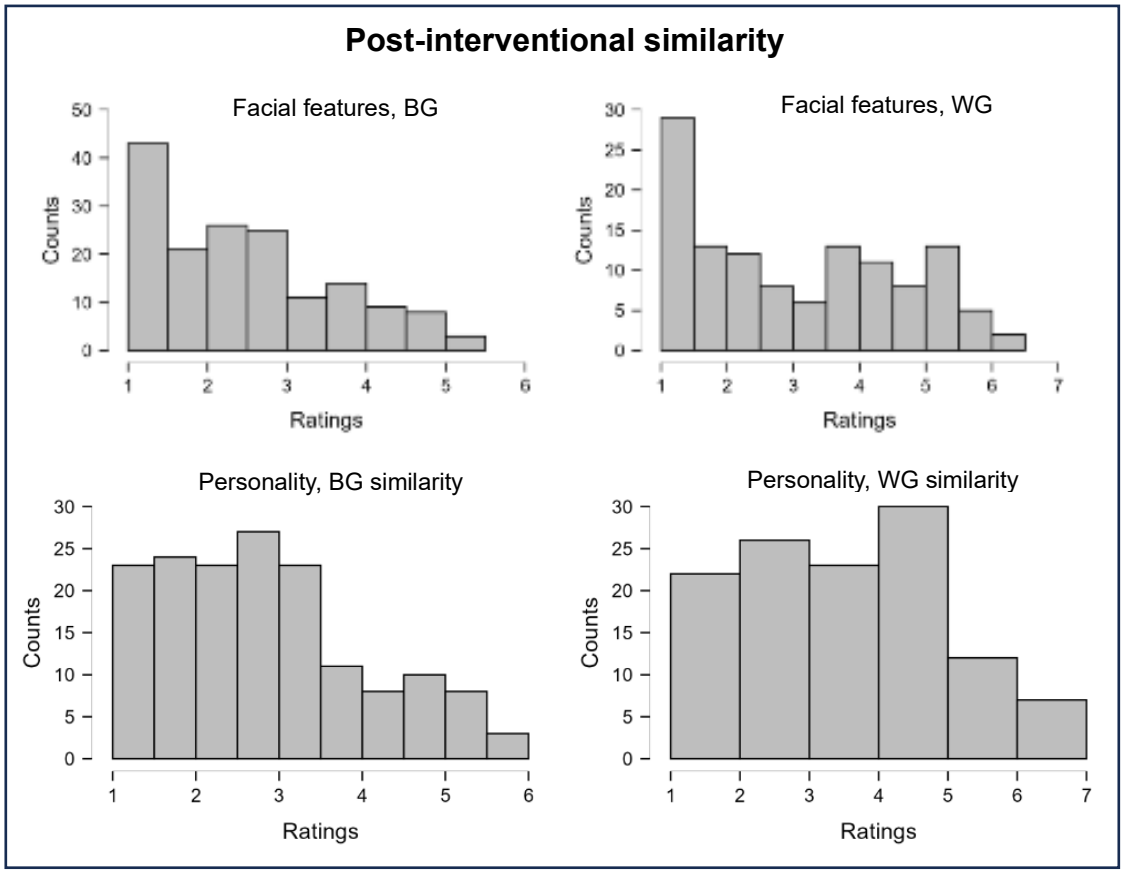


Figure 8: Similarity ratings of the eight students in physical facial features were right-skewed in both similarity types; WG and BG Similarity ratings of the two other traits (personality and social life) were also skewed but not as strongly.

Table 4: Results of the Shapiro-Wilk test conducted on post-interventional WG and BG similarity ratings of the eight students over the three traits measured.

Task Dimension	Similarity type	Sample size	W	p-value
	WG	120	0.928	< .001

Physical facial features	BG	160	0.941	< .001
Personality	WG	120	0.964	.003
	BG	160	0.955	< .001
Social life	WG	120	0.968	.006
	BG	160	0.969	.001

This showed that all the ratings from the three tasks were not normally distributed. We conducted three linear mixed models to test hypothesis 1 over the three traits measured. For each trait, we conducted a linear mixed model with ratings as a dependent variable and one fixed effects factor (similarity type: within-group similarity vs. between-group similarity). Subject ID was added as a random effect.

3.4.3 Results

Linear mixed effects models showed a significant main effect of similarity type in the dimensions tested $F(1,14.7)=12.1$, $p=0.003$; $F(1,9)=7.59$ $p=0.022$; $F(1,9)=6.12$, $p=0.035$ for physical facial features, personality, and social life. Within-group similarity ratings were consistently higher than Between-group similarity ratings in all three dimensions we tested. Higher within-group similarity than between-group similarity indicates that participants categorized the students into two groups after the intervention. Participants rated students who studied the same study field (whether they are AH or NAT students) to have similar personalities and social lives, while they rated students who studied different study fields to be less similar to each other. Surprisingly, participants also reported social categorization over the physical facial features dimension. This indicates that participants reported that Natural Science students looked physically similar to each other and Arts and Humanities students looked like each other. Unfortunately, we couldn't test whether this was also the case before the intervention or not. So, we couldn't conclude that this pattern of social categorization is only the result of our intervention and not caused by an actual similarity inherent in our stimuli.

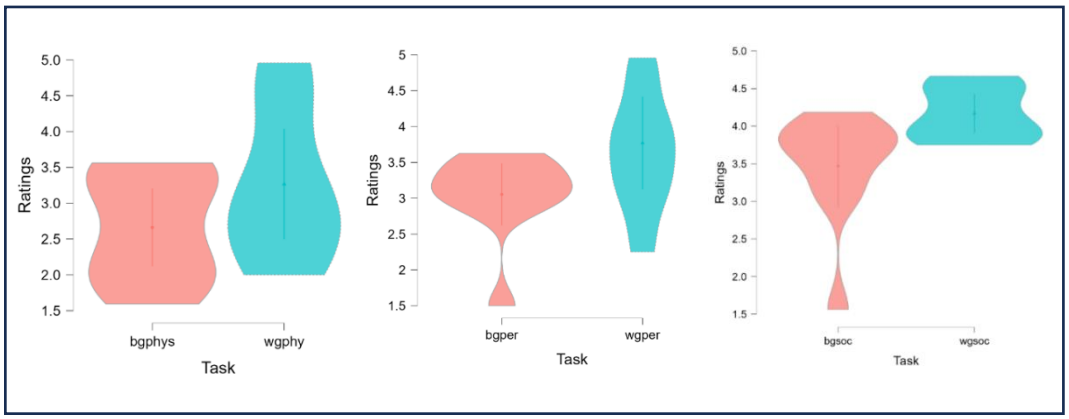


Figure 9: Post-interventional similarity ratings in physical facial features a), personality b), and social life c). After knowing the group membership of the eight students. Participants reported that the similarity between AH and NAT students was significantly lower than the average within-group similarity in AH and NAT groups. Meaning that they perceived students of each group to have similar facial features, personality, and social life and both groups to be different from each other.

3.5 Main Experiment

3.5.1 Participants

We conducted a behavioral experiment over two days in Bonn EconLab, Bonn Germany (Bonn EconLab, 2023). 23 Bachelor students 15 NAT and eight AH students. The participants included 11 males and 12 females with an average age of 22.39 years old. Notably, seven participants indicated that they considered themselves as members of minority groups.

3.5.2 Study Design

The experiment was conducted on computers using PsychoPy software and followed a within-subject design. The following tasks were included in the experiment, group identification questionnaire, pairwise similarity task, and Meet the Students Task. The experiment flow was as follows; first, participants were received at the EconLab and were instructed briefly about the experiment. Then the participants conducted the Group Identification Questionnaire, followed by the Pairwise Similarity Task. After that, we introduced the students to the participants using the Meet the Students Task and tested their memory using the Group Membership Recall task. Finally, the participants repeated the Pairwise Similarity Task.

3.5.3 Data Analysis

As was done in the pilot study, between-group similarity ratings and within-group similarity ratings were separated, and we pooled each type of similarity ratings together. We

collected data from 23 participants, each one reported 16 between-group similarity ratings and 12 within-group ratings for each pairwise similarity task. Since we performed the task before and after the intervention, there were in total six tasks per subject. So, for 23 subjects, we had eventually 368 between-group ratings and 276 within-group ratings. We show a descriptive table of the data in Table 5. First, we tested for data normality using the Shapiro-Wilk test. It showed all data groups significantly deviated from normality.

Table 5: Descriptive statistics of all the six pairwise similarity tasks conducted (3 traits x 2 time points) in addition to the results of the Shapiro-Wilk test on each group of results (within-group ratings (WG) vs. Between group ratings (BG)). BG ratings are by nature more than WG ratings, resulting in an unbalanced design. Shapiro-Wilk test was significant for all data groups meaning that no data group was normally distributed.

Dimension	Time	Similarity type	Sample size	Mean	SD	W Statistic	p-value
Physical facial features	Pre	WG	276	3.58	1.5	0.96	<.001
		BG	368	3.19	1.3	0.97	<.001
	Post	WG	276	3.77	1.5	0.95	<.001
		BG	368	3.36	1.4	0.96	<.001
Personality	Pre	WG	276	3.99	1.3	0.97	<.001
		BG	368	3.94	1.2	0.98	<.001
	Post	WG	276	4.3	1.3	0.97	<.001
		BG	368	3.4	1.3	0.95	<.001
Social life	Pre	WG	276	4.1	1.3	0.97	<.001
		BG	368	3.98	1.3	0.97	<.001
	Post	WG	276	4.23	1.3	0.97	<.001
		BG	368	3.49	1.3	0.97	<.001

Then, we conducted a repeated measures ANOVA to test whether the pairwise similarity ratings changed from before to after the intervention. We created one separate model for each of the three dimensions we tested, physical facial features, personality, and social life.

3.5.3.1 Dissimilarity (Distance) Values

In order to convert the 28 similarity ratings for each pairwise similarity task into a visual map, we first converted the similarity value into distance values. We did that by reverse coding the similarity value where a similarity of “7” becomes a dissimilarity of “1” and a

similarity of “1” becomes a dissimilarity of “7”. The dissimilarity values were then used as Euclidean distance to draw a map of how distant each person is perceived to be in comparison to another. However, since we have eight persons, any dissimilarity matrix D requires a seven-dimensional Euclidean space to be visualized perfectly. So, in order to visualize the dissimilarity matrix geometrically in only two-dimensions, we used a classical multi-dimensional scaling (MDS) implemented in MATLAB function “cmdscale” . cmdscale function generates six-dimensional distance matrices, the number of dimensions in the generated matrix equals the rank of matrix B, where B is the inner product matrix of the input dissimilarity matrix (Gower, 1966; Torgerson, 1952, 1959). This number of dimensions is also the minimum number of dimensions in which the dissimilarity distances could be represented. We then used dissimilarity values from the first two dimensions to draw an approximation of the dissimilarity maps in two-dimensional spaces. For more information about MDS calculation, see (Seber, 1984).

3.5.4 Results

3.5.4.1 Hypothesis 1

Physical Facial Features: The repeated measures ANOVAs for physical facial features showed significant main effects for Time (pre- vs. post-intervention) ($F(1,275) = 6.25$ $p= 0.013$) and for Measurement type (Within-group vs Between-group similarity ratings)(type $F(1,275)= 45.86$ $p<.001$) but no significant effect of the interaction between Time and Measurement. Post-hoc t-tests showed that within-group and between-group similarity ratings did not significantly change from before to after the intervention ($t = -1.65$, $p_{Holm} = 0.106$; $t = -1.95$, $p_{Holm} = 0.106$).

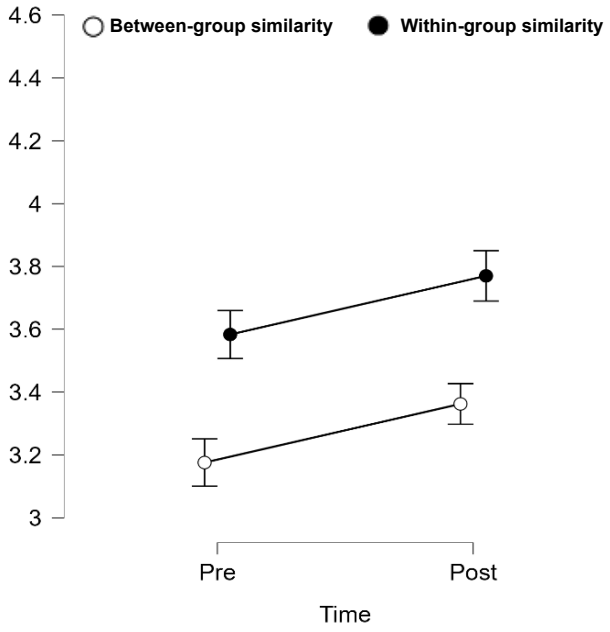


Figure 10: shows the average similarity ratings in Physical facial features of the eight newly encountered students whom the participant viewed. Before knowing the group membership of the students, there was no significant difference between WG and BG similarity ratings. This didn't change after the intervention as WG and BG didn't change significantly from before to after the intervention.

Personality: The repeated measures ANOVAs for personality showed significant main effects for Measurement type ($F(1,275)= 55.26$, $p<.001$) and Time*Measurement type interaction ($f(1,275) = 43.8$, $p<.001$). Post-hoc t-tests showed that within-group similarity ratings increased significantly from before to after the intervention ($t = -3.168$, $p_{Holm} = 0.003$) while between-group similarity ratings decreased significantly ($t=4.888$, $p_{Holm}<.001$).

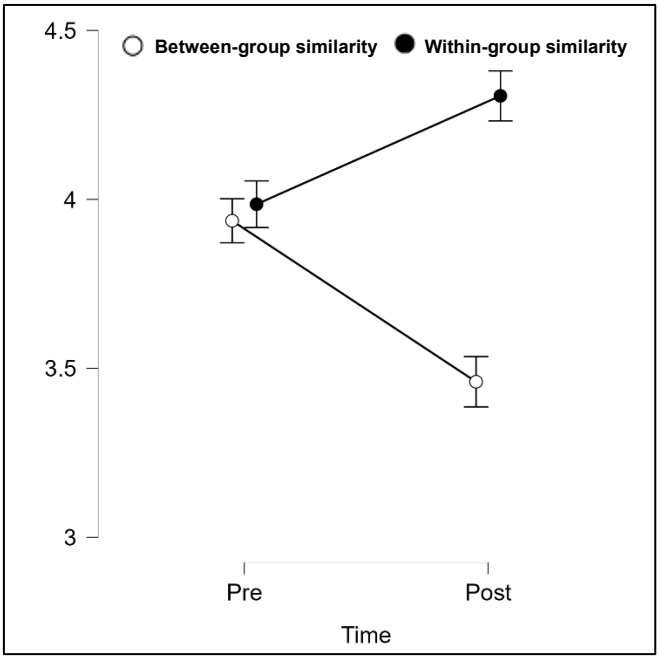


Figure 11: Shows the average similarity ratings in personality of the eight newly encountered students whom the participant viewed. Before knowing the group membership of the students (Pre) the WG and BG similarity ratings were very similar. So, no social groups were formed, i.e. no clustering of students. However, after the intervention, there was a significant decrease in perceived BG similarity p_{Holm}<.001 that distinguished the two groups from each other, and a significant increase in WG similarity ratings p_{Holm}=003 which resulted in clustering of each group.

Social Life: The repeated measures ANOVA for social life showed significant main effects for Time ($F_{(1, 275)}$ = 5.718, p=0.017), Measurement type ($F_{(1, 275)}$ = 33.85, p<.001), and Time*Measurement type interaction ($F_{(1, 275)}$ = 17.716, p<.001). Post-hoc t-tests showed no significant change in within-group similarity ratings from before to after the intervention (t = -1.272, p_{Holm} = 0.409) while between-group similarity ratings decreased significantly (t=4.863, p_{Holm}<.001).

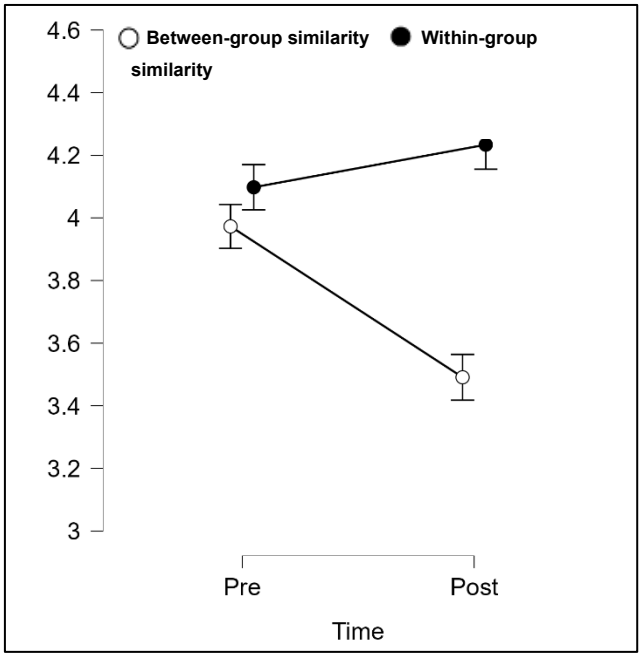


Figure 12: shows the average similarity ratings in social life of the eight newly encountered students whom the participant viewed. Before knowing the group membership of the students (Pre), the WG and BG similarity ratings were very similar. So, no social groups were formed, i.e., no clustering of students. However, after the intervention, there was a significant decrease in perceived BG similarity, pHolm<.001, that distinguished the two groups from each other and a slight increase in WG similarity ratings, which also helped the clustering of each group.

3.5.4.2 Mapping the Perceived Group Topology

The MDS we conducted generated six dissimilarity maps, one for each combination of time point (pre- vs. post-intervention) and task (similarity in physical facial features, personality, or social life). When comparing the pre-intervention dissimilarity maps of personality and social life to post-intervention maps, we noticed a striking change in perceived group topology. As presented in Figure 13, there was an obvious reorganization of the dissimilarity maps in personality and social life after revealing the group membership of the students. Since the students only talked about their group membership, we can assume that the content of the videos was the cause to the change in perceived group topology. Social Categorization is shown in post-interventional dissimilarity maps of personality and social life. This can be viewed as clustering (small dissimilarity) between members of the same social group and large dissimilarity in ratings between students who belong to different social groups.

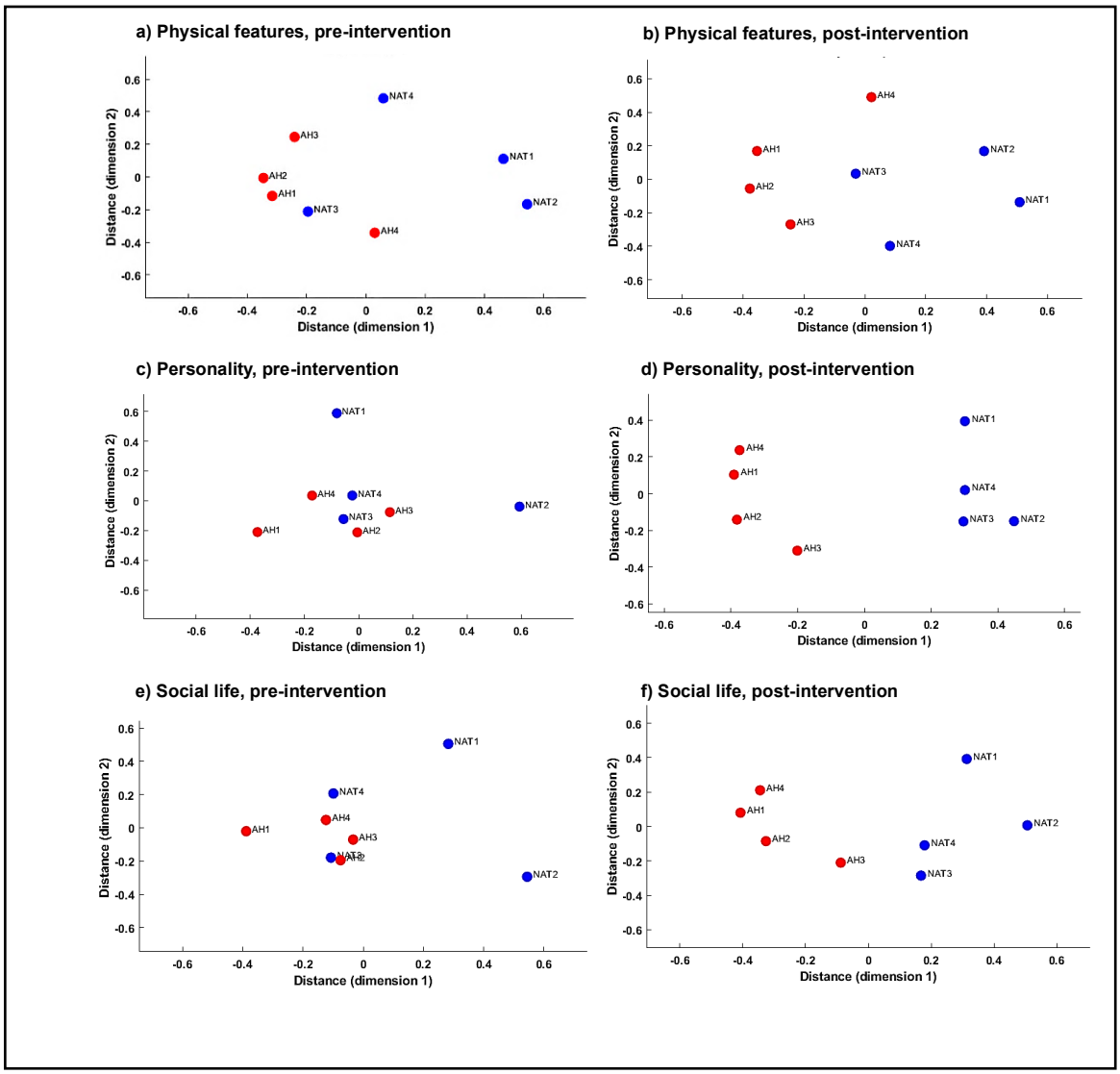


Figure 13: Visualization of the dissimilarity maps in physical facial features, personality, and social life of the eight students before and after the intervention. The maps were generated using classical multidimensional scaling, so they are approximations of the actual maps. Each dot represents one student where blue dots are the four Natural Science students and the red dots are the four Arts and Humanities students. Here the dissimilarity is represented as Euclidean distance so a larger distance between any two points means that these two students were perceived to be dissimilar from each other. The maps show clear social categorization patterns (small WG distances and large BG distances) in similarity ratings of personality and social life after the intervention which wasn't there before the intervention. Dimensions 1 and 2 represent the two spatial dimensions X and Y (distances between Likert-scale based ratings) see 3.5.3.1

3.6 Discussion

This study investigated how individuals perceive the similarity between persons in the context of social categorization. Rather than focusing on group-level constructs such as homogeneity, the research centered on person-level similarity judgments influenced by the presence of group membership information. Specifically, it asked: How does knowing someone's academic group (Natural Sciences vs. Arts & Humanities) influence

participants' judgments of how similar two people are to one another? We used the pairwise similarity task as an indirect method to construct a map of the social group's topology. The main difference between this task and previous measures is that participants are asked about individual group members in comparison to other group members. This avoids too general judgments of groups using the group-based tasks and avoids rating of individuals in an abstract way as what was done in RGM task we used in study 1. The results demonstrated that group information significantly shaped perception of group members. Participants perceived greater similarity between individuals who belonged to the same academic group after knowing the group membership of the individuals. This aligns with theoretical accounts of social categorization as a cognitive clustering mechanism (Tajfel & Turner, 1979)

3.6.1 Interpreting the Influence of Group Cues

By employing a pairwise similarity task, this study moved beyond abstract group-level constructs to quantify how people compare individuals to each other making it possible to construct perceived similarity maps along different dimensions. Importantly, the participants had minimal information about the persons beyond the face images and their academic group. This limited exposure underscored the power of our intervention and established the intervention to be causal in the social categorization process.

3.6.2 Relevance to Neuroscientific Models of Person Perception

The study's design is particularly valuable as it lays groundwork for future neuroimaging research aimed at mapping perceived person-to-person similarity onto brain activity patterns. Rather than assuming categorical processing at the group level, this approach allows researchers to examine representational similarity between individuals as encoded in neural networks. This opens new possibilities for identifying the brain mechanisms underlying social judgments, stereotypes, and even discriminative behavior.

3.6.3 Methodological Strengths and Constraints

This study's key strength was the careful control of stimulus features: facial images were standardized (grayscale, cropped, luminance-matched), and participants were prevented from over-individuating stimuli by limiting extraneous information. This helped to isolate the impact of group membership information on perceived similarity without confounding perceptual variability. Still, there were several limitations that deserve consideration. The

binary group manipulation (Natural Sciences vs. Arts & Humanities) may not capture the complexity of real-world group dynamics or multidimensional similarity judgments. Additionally, the reliance on visual stimuli, even in a person-perception task, may have triggered visual-based heuristics that partly confound cognitive-level similarity judgments. Future iterations might include multimodal stimuli or to represent persons. Despite the success of showing social categorization, the pairwise similarity task (as with the Similarity rating task (SIM)) was sensitive to the trait or dimension that was rated. We cannot expect that participants perceived NAT and AH students to be different in all behavioral and physical traits. For example, there's no reason to think that NAT students had the same height or to physically look like each other. Therefore, the pairwise similarity rating results are only indicative of the perceived similarity among group members along the specific dimension in question.

3.6.4 Implications and Future Directions

The results highlighted the possibility of mapping the fine-grained differences in perception of social group members and emphasized that category labels can reshape interpersonal judgments. Methodologically, the study introduces a behavioral paradigm that can serve as a bridge between psychological measures and representational similarity analysis (RSA) in neuroscience experiments. This direction moves away from asking which regions are activated by group membership per se, and instead asks whether brain activity patterns reflect perceived psychological distance between individuals. If successful, this could lead to more precise models of social information processing, integrating behavioral and neural similarity metrics in novel ways (Weaverdyck et al., 2020; Popal et al., 2020).

3.6.5 Future research

The current research could be expanded in many directions. For example, a) Testing the paradigm across different social categories (e.g., gender, political identity) to assess the robustness of category-based similarity effects. b) Examining the Individual differences in how cognitive style, group identification, or ideological orientation moderate the social categorization effect. c) Using this task in fMRI behavioral similarity maps with neural representational geometry.

4 Study 3: Neural Correlates of Social Categorization

4.1 Introduction

Neuroimaging experiments with an interest in social category perception started as early as the beginning of the field of functional MRI itself. The early experiments focused specifically on racial outgroup effects between White Caucasian and Black African races. (Hart et al., 2000). This was motivated by understanding social discrimination based on race, which plagued the USA. The pioneering study to link a behavioral measure with neural responses to racial outgroup faces was conducted by Phelps et al., 2000. In this experiment, White participants completed an Implicit Association Test (IAT) that assessed unconscious racial biases, which was then correlated with fMRI data collected while participants viewed photographs of unfamiliar Black and White faces. The researchers found that greater amygdala activation in response to Black faces, relative to White faces, was positively associated with stronger pro-White/anti-Black bias on the IAT. This finding provided early evidence that implicit racial attitudes are reflected in neural responses in regions associated with emotion and threat processing, like the Amygdala.

Extending this line of research, Cunningham et al. (2004) presented participants with Black and White faces under two timing conditions: a subliminal condition (30 milliseconds) and a supraliminal condition (525 milliseconds). During the brief, subliminal exposure, amygdala activation was significantly greater for Black faces than for White faces, replicating and refining the findings of Phelps et al. (2000). However, in the supraliminal condition, this amygdala difference diminished. Instead, greater activation emerged in regions implicated in cognitive control, including the dorsolateral prefrontal cortex (DLPFC) and anterior cingulate cortex (ACC). These results suggest that while initial, automatic responses to outgroup faces may engage the amygdala, more controlled, deliberative processing, as allowed by longer exposure, may recruit cortical mechanisms that regulate or override initial biases.

Together, these studies underscore the temporal sensitivity of neural responses to race and highlight a dual-process framework in which automatic and controlled processes interact to shape intergroup perception. So, the simple story that was previously suggested that White people are threatened by Black faces might not be exactly correct, and amygdala activity in the earlier studies might be due to another mental process.

Overall, the experiments conducted by Phelps et al. (2000) and Cunningham et al. (2004) showed the importance of associating a properly curated behavioral measure with neuroimaging in such experiments. However, the first published experiment that approached neuroimaging of outgroups differently came only in 2008.

4.1.1 Racial Versus Outgroup Effects

Van Bavel et al., (2008) used a mixed group paradigm where they created two groups of people, each containing White and Black people, and gave an arbitrary label to each group, one being tigers and the other being leopards. Then they allocated participants randomly to either one of these minimal groups. This ingenious design allowed them to test ingroup vs. outgroup effects separately from White vs. Black races. For example, Van Bavel et al. found that viewing new in-group members' faces was associated with greater activity in bilateral fusiform gyri than viewing out-group members' faces. Although this effect was reported before (Golby et al., 2007; Lieberman et al., 2005), it was attributed to more familiarity with the racial in-group, an effect reported in earlier work (Golby et al., 2001). This is because the previous experiments confounded race with social groups, where ingroups were always White Caucasian subjects.

Van Bavel and his colleagues' work showed that faces of novel ingroup members elicited greater activity in the bilateral fusiform gyri and the left OFC, amygdala, and ITG (Van Bavel et al., 2008). In a follow-up experiment, the FFA showed greater activation to ingroup faces compared to outgroup and control faces, and that the increased FFA activity to ingroup faces was positively correlated with better recognition memory for ingroup versus outgroup faces (Van Bavel et al., 2011). The results from both experiments were independent of the races of the presented faces. Hence, they emphasized the necessity of distinguishing between neural mechanisms underlying the perception of faces from other races, also known as the "Other Race Effect", and those involved in perceiving social groups. The Other Race Effect refers to the robust psychological phenomenon in which individuals have worse memory for faces of other races compared to faces of their own race (Meissner & Brigham, 2001). Van Bavel and colleagues demonstrated that social categorization can override racial categorization, suggesting that neural responses to race are not fixed but are modulated by dynamic group membership. Some later studies started

to use a mixed group similar to van Bavel et al. 2008 (Contreras-Huerta et al., 2014; Van Bavel et al., 2011; Yan et al., 2019).

4.1.2 Univariate Analysis

Early neuroimaging experiments used a simple logic in their analysis. They tried to find which brain regions show increased blood-oxygen-level-dependent (BOLD) signal when participants thought of, viewed, or interacted with outgroup members in comparison to ingroup members. A plethora of interesting findings were reported that include different aspects of social activities. For example, outgroup faces were more likely to induce higher activity in brain regions like the anterior cingulate cortex (ACC) and dorsolateral prefrontal cortex (DLPFC), which was suggested to be related to top-down control over pre-existing stereotypes towards outgroup members (Amodio, 2014; Kubota et al., 2012).

Another line of research focused on empathy and emotion recognition in outgroup vs ingroup members, which included participants viewing ingroup or outgroup members being exposed to painful stimuli. For example, Hein et al., (2010) found – using a minimal group paradigm – a stronger brain activation in the left anterior insula cortex when participants viewed an in-group member in pain in comparison to an out-group member. This left insular activity correlated with higher frequency of costly helping of ingroup members and higher scores on the Empathic Concern Scale (Hein et al., 2010). Differential emotion-recognition and empathy for ingroup vs outgroup members could be attributed to intergroup threat and consequently result in intergroup violence (Lantos & Molenberghs, 2021). In addition to the previous findings, researchers studied the univariate difference in brain activity between ingroup and outgroup members in tasks like social categorization (Feng et al., 2011), impression formation (Li et al., 2016) and dehumanization (Bruneau & Saxe, 2018). A recent meta-analysis examined and compiled the univariate analysis results across different tasks and grouping paradigms, and found consistent differential activity depending on the social group associated with the task mainly in the prefrontal cortex, fusiform gyri and the Insula (Merritt et al., 2021).

4.1.3 Multivariate Analysis

Univariate functional magnetic resonance imaging (fMRI) analysis typically examines activation in different voxels separately, testing whether individual voxels show statistically significant signal changes in response to specific experimental conditions. While this

approach has been instrumental in mapping functional brain regions, it has the following notable limitations. a) Loss of fine-grained information: Because univariate methods focus on mean signal differences across conditions, they fail to capture distributed patterns of activation across multiple voxels. b) Inability to detect representational content: Univariate analyses primarily identify whether a brain region is more active in one condition than another, but do not reveal how information is represented within that region. c) Problematic assumptions of spatial covariance: In univariate analysis, each voxel is modeled independently. Therefore, univariate approaches ignore spatial covariance, potentially missing important distributed coding mechanisms (Kriegeskorte et al., 2006).

Multivariate pattern analysis (MVPA) addresses these limitations by analyzing patterns of activity across multiple voxels. Instead of averaging signals, MVPA examines how different conditions evoke distinct spatial patterns of activation, allowing researchers to infer what type of information is represented in a given region (Carlson, 1998; Haxby et al., 2001; Nili et al., 2014; Oosterhof et al., 2016). One key advantage of this approach is that it allows enhanced sensitivity to information representation: unlike univariate methods, which focus on mean activation differences, MVPA detects fine-grained, distributed patterns of activity, making it possible to distinguish between neural representations of different stimuli even in regions with overlapping activation (Kriegeskorte et al., 2006; Kriegeskorte & Bandettini, 2007).

4.1.4 Decoding Neural Representations

MVPA allows researchers to decode stimulus categories from neural activity patterns. For example, Haxby et al., 2001 demonstrated that face and object representations in the ventral temporal cortex are distributed rather than localized to specific brain regions. This study showed that even within face-selective regions, there is information about other stimulus categories, challenging the assumption of category-specific regions.

4.1.5 Representational Similarity Analysis (RSA)

RSA is an analysis method used in different scientific domains, and it has recently been adopted in neuroscience methods, as it has been shown to be beneficial in multivariate pattern analysis (Kriegeskorte et al., 2008; Popal et al., 2020). RSA is a high-level analysis that builds on primary findings of similarity between stimuli or similarity between responses to the stimuli. For example, an experimenter can ask a participant to indicate the degree

of similarity in function between 3 items, e.g., muscles, bones, and kidneys, and also the degree of similarity in lifestyle between professionals working on these items, i.e., orthopedist, physical therapist, and nephrologist. These two questions result in two similarity patterns, one for the item function and one for the lifestyle of the professionals. RSA can then be used to compare the similarity patterns. In the case described here, we expect that the two similarity patterns correlate positively with each other.

4.1.6 RSA in Multivariate fMRI analysis

Since in MVPA, the data is collected from a group of brain voxels at once, informative brain regions are defined by moving a spherical or circular multivariate “searchlight” (Kriegeskorte et al., 2006). A surface-only searchlight of 10-mm radius, for example, will comprise of around 180 surface nodes when laid on fsaverage6 template which contains 40962 nodes per hemisphere (Fischl, 2012). When conducting a statistical test, or calculating a measure, the convention is to map or view the test statistic or measure on the center node or voxel of the searchlight (Kriegeskorte et al., 2006). In multivariate fMRI analysis, a Representational Dissimilarity Matrix (RDM) is a correlation matrix based on all pairwise comparisons between the BOLD signal responses of group of voxels evoked by a set of stimuli. The group of voxel can be defined based on a functional cluster, anatomical region, or local neighbors (as in a searchlight analysis). RSA then consists in comparing an RDM from a brain region, e.g., a disc in the right anterior temporal lobe (Rt. ATL) with either a) another fMRI signal RDM, e.g., from left ATL, b) a behavioral RDM, e.g., from a Pairwise Similarity Task, or c) a manually constructed RDM which represents a specific theoretical model. RSA is a flexible tool which allows the comparison between representations from different domains, e.g., data from behavioral measures or data from neurological measures e.g., fMRI, EEG, fNIRS, etc., and even between species (monkey vs. human) and at different spatial scales (single neuronal activations vs regional activations) (Kriegeskorte, Mur, Ruff, et al., 2008; Ritchie et al., 2021).

4.1.7 Linear Discriminant Contrast (LDC) and t-value (LDt)

Different methods were proposed and used to measure the dissimilarity between stimuli, each dissimilarity measure is different and thus has an impact on the resulting RDMs. The most common measures used in the literature are reversed Pearson’s correlation (1-r, where r is Pearson’s correlation coefficient) and Mahalanobis distance. More

sophisticated measures derived from the Mahalanobis distance that account for both variance and covariance structure in the data by estimating a noise ceiling are called Linear Discriminant Contrast (LDC) and Linear Discriminant t-value (LDt) (Popal et al., 2020). The cross-validated Mahalanobis distance is called Linear Discriminant contrast (LDC).

Walther and his colleagues (Walther et al., 2016) discussed the different types of dissimilarity measures and through simulated data showed that a cross-validated Mahalanobis distance is the most reliable measure of dissimilarity especially when the distances between data points are large. This is because, with increasing Euclidian distance between two conditions, the noise level also increases. It has been suggested to normalize the LDC by an estimate of its standard error, which results in linear-discriminant t-value (LDt) and can be used as an inferential measure of stimulus dissimilarity. Overall, LDt seems to be a robust multivariate measure of the dissimilarity between two vectors, each one representing the state of all searchlight voxels in response to an experimental condition, for example, the dissimilarity between response to an ingroup face compared to an outgroup face (Nili et al., 2014; Walther et al., 2016).

4.1.8 Anatomical Versus Functional Brain Alignment (Hyperalignment)

Since we recruit multiple participants in our studies, we conduct group-level analysis to find activation patterns that generalize to the tested sample, and by inference, to the population from which the sample came. To perform voxel-wise group tests, it's a necessity to align individual brain scans of our participants into a common space. The most common method consists in aligning the brains based on anatomical structures, a method called anatomical alignment or anatomical normalization. This method assumes that after normalization, a given voxel is associated with a given function in all the brains included in the sample. However, it has been shown that people show individual differences in anatomical and functional regions of the brain, which can lead to high variability of responses across subjects, even in response to identical stimuli (Andreella et al., 2022).

Hyperalignment (functional alignment or functional normalization) is an alternative method of aligning subjects' brains with each other in group-level analysis. It has been proposed to align shared functional information into a common space (Haxby et al., 2011).

In order to perform hyperalignment, participants must be exposed to the same stimulus sequence throughout a scan, such that one can assume that the sequence of stimulus processing and mental states of the participants is also similar during the scan. For example, we assume that they will have similar emotions or cognitive experiences in response to a long visual stimulus, such as a movie or a part thereof. Then, a procrustean transformation is applied to the data in order to find a common space where the information from different individual brains is shared. Hyperalignment dramatically improves between-subject decoding (Guntupalli et al., 2016; Haxby et al., 2011). However, since the whole-brain implementation of this approach remixes data from relatively distant voxels, one cannot analyze or interpret the data in the common functional space voxel by voxel. To enable localized analyses, the analysis is usually done iteratively in a series of regions of interest (ROIs), such as a searchlight sphere or disc (Haxby et al., 2020).

4.1.9 Improved Understanding of Face Perception Mechanisms

Multivariate analyses have revealed key insights into face perception, such as the fact that the fusiform face area (FFA) contains distinct activation patterns for individual face identities, which univariate methods failed to detect (Goesaert & Beeck, 2013; Kim et al., 2019; Nestor et al., 2011; Verosky et al., 2013; Weiner & Grill-Spector, 2010). The anterior temporal lobe (ATL) plays a crucial role in encoding conceptual knowledge about faces, supporting the idea that face processing is not confined to the occipitotemporal cortex (Anzellotti et al., 2014; Yang et al., 2016). Studies using representational similarity analysis (RSA) have shown that facial identity representations are distributed across multiple regions, reinforcing the idea of a networked rather than modular system (Collins et al., 2016; Kriegeskorte et al., 2007).

4.1.10 Brain Responses to Different Social Groups

What we know so far is that overall, stimuli associated with individuals from different social groups induce different brain activity in various brain regions. However, current literature doesn’t shed light on how social group structure (similarities and differences between individuals) is encoded and represented in the brain. In addition, since social group perception changes according to the trait in question (for example, trait prototypicality), there’s currently no method for measuring the change in perception of individuals belonging to the social groups based on the change of the trait in question. In other words,

we don't know how the perceived similarity between members influences the neural representation of these individuals.

4.2 Aim of the Project

Find the brain regions or networks that change their neural representation of persons in correlation with the behaviorally reported change in perception of these persons.

4.3 Pre-Registration

Before data collection, we preregistered the study on the Open Science Framework, here: https://osf.io/q6du8.

4.4 Hypotheses

1. Revealing the group membership of a set of newly learned students will lead to an increase in *perceived within-group similarity* and a decrease in *perceived between-group similarity*.
2. In at least one brain region, revealing the group membership of presented faces will lead to increased similarity between the neural representations elicited by faces that belong to the same group and decreased similarity between representations elicited by faces that belong to different groups.
3. In at least one brain region, average pairwise similarity between the neural activities elicited by presentation of the faces of our learned students will be positively correlated with average ratings of pairwise similarities in personality and social life of the respective students.

4.5 Study Design

In this experiment, we use the same behavioral experiment design as in Study 2: we measured the perceived pairwise similarity between the eight students along physical facial features, personality, and social life. We used a within-subject design, and collected pre-intervention ratings, performed the intervention, then collected post-intervention ratings. In addition, participants performed an fMRI task before and after the intervention. The intervention consisted in revealing the group membership of each of the eight students. In the fMRI task, participants were presented with the eight faces and answered a yes-no question about the group membership of the face they were looking at.

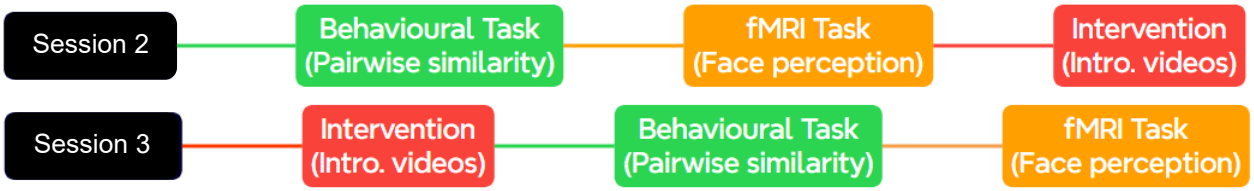


Figure 14: Experiment Workflow in the second and third sessions.

4.5.1 Recruitment

Participants were recruited via the EconLab mailing list and by public advertisements. The interested subjects first filled in an online Qualtrics questionnaire. The group identification questionnaire was added as an online questionnaire with the same questions used in Study1 and Study2, but in this experiment, the participants answered this questionnaire online before inviting them to the experiment. However, the degree of group identification did not play a role in participants' inclusion. We applied the following exclusion criteria:

- Mixed group identity: participants who identify as both NAT and AH students.
- MRI contraindications: Body implants, metal accessories, claustrophobia, and big tattoos.
- Previous familiarity with any of the eight students included in our stimuli.
- If they had watched the movie The Grand Budapest Hotel (2014) before.
- Left-handedness or age below 18 or above 30 years old.

After applying the exclusion criteria, we invited 20 right-handed native German, bachelor, or master students (females = 10) to participate in the experiment.

4.5.2 Stimuli

Most of the stimuli were the same as Study 2 stimuli. We used the a) Introductory Videos, b) Color Portraits, and c) Cropped Greyscale Faces. In addition, we generated face clips for the fMRI task as follows.

4.5.2.1 Face Clips

Four random 0.5-second clips were extracted from each of the introductory videos. Because each video originally contained 30 frames per second, our half-second clips contained 15 frames each. We converted each frame into a picture and applied the same preprocessing as we did to our Cropped Greyscale Faces. Then we recreated the face clips from the pre-processed frames. We included different expressions in the videos. And,

at least, one of the four clips contained a clip showing the student blinking. The clips did not have any sound.

4.5.3 Tasks

The experiment has two types of tasks organized in six time blocks, four of which are conducted on a computer screen and two are conducted inside an MRI scanner. We used the same tasks as in Study 2, namely: a) Pairwise Similarity Task, b) Intervention (Meet the Students Task), c) Group membership recall Task, d) Group Similarity task (SIM), *e) Percentage estimate task (PER)*. In addition, we added a *Face Perception Task as follows:*

4.5.3.1 Face Perception Task

This task was conducted in an MRI scanner. It was comprised of six runs; each one was divided into 56 trials. In each run, we first told participants that they had to answer one main question about the faces they viewed. The question was “Does the person you just saw belong to the same study field as the person before?”. Each trial consisted of three 500-ms presentations of the face separated by 50-ms black screens (Castello et al., 2021) then followed by a three-second fixation cross, where participants are required to answer the main question. Stimuli are presented in a one-back DeBruijn cycle order (Aguirre et al., 2011).

4.5.3.2 PC Movie Task

participants watched the first 50 minutes of the film The Grand Budapest Hotel (2014) dubbed in German.

4.5.3.3 Volume Adjustment

his task was conducted only in the first scanning session; the participant laid in the scanner and repeated watching the last five minutes of the first half of the movie. We instructed the participant to gesture to us using their right hand either to increase or decrease the volume or to keep it levelled. During this time, the experimenter’s assistant stayed in the scanning room while the participant was being scanned and conveyed the requests of volume change to experimenter. After the scanning sequence was finished, we asked the participant if the headset and volume setup were comfortable and if the movie sounds were clear. Then, we adjusted the volume further if needed.

4.5.3.4 MRI Movie Task

We presented the next 50 minutes of the same film until the end of the film. The 50 minutes were divided into 5 sections of different lengths in order not to cut film scenes or conversations as in (Castello et al 2020).

4.5.3.5 Fieldmap Scan

This was a short two-minute task where the participant laid still in the scanner. No task was required from the participant during the scan. The purpose of this scan was to measure the inhomogeneities in the magnetic field of the scanner. These inhomogeneities specifically lead to distortions of the functional scans especially in skull regions that contain air close to the brain, most commonly in frontal and temporal poles.

4.6 Experiment Flow

4.6.1 First Session

Once arrived at the facility, participants were welcomed and signed documentation related to their consent to participate in the study, data storage, and MRI safety information. Then, they conducted the PC Movie Task on a PC. After that, the participants were led to the MRI room. The participant wore MRI-compatible earphones and conducted the Volume Adjustment while we acquired the structural scan. After that, the participants conducted five runs of the MRI Movie Task. After a short rest, the participants briefly exited the scanner, swapped their headphones for earplugs, reentered the scanner and conducted four runs of the Dynamic Face Localizer Task.

4.6.2 Second Session

On a PC, the participants conducted the Pairwise Similarity Task along three traits (Physical face features, personality, and social life). Next, the participants conducted six runs of the Face Perception Task, followed by a fieldmap scan, then exited the scanner. After a short break, the participants conducted the Meet the Students Task, where they watched the introductory videos on a PC.

4.6.3 Third Session

The participant started the session with the Meet the Students Task once again, followed by the Group Membership Recall Task. Then the participants repeated the Pairwise Similarity Task over the same three traits as was done before the intervention. After that,

the participants laid in the scanner and conducted six runs of the Face Perception Task followed by a fieldmap scan.

4.7 Data Acquisition

Before conducting the main experiment, we piloted the experiment with three German-speaking master students. These participants conducted parts of the experiment in order to test our experimental setup and the analysis code.

For the main experiment, the number of required participants was calculated based on a power analysis conducted using GPower software (Faul et al., 2007) with an alpha criterion of 0.05 and an effect size of 0.496. The fMRI study was conducted at the MRI Core Facility (Life & Brain Research Center) of the Medical Faculty of the University of Bonn. For each participant, the study was divided into three sessions, each session lasting two to two and a half hours, with 2-4 days between each session and the next. The participants filled out a consent form explaining the experiment sessions, the risks and contraindications of MRI scanning, and our data privacy policy. In addition, each time the participant entered the MRI scanner room, they filled out a consent form where they acknowledged the absence of an MRI contraindication. All MRI images are collected using a Siemens 3T Magnetom TRIO. In the functional scans, the participants were provided with earplugs to protect their hearing from the loud noises of the MRI scanner. Only during the MRI Movie Task, the participants wore MRI-compatible earphones which dampened the scanner noise by 30dB. While inside the scanner, participants viewed the instructions and stimuli via a mirror system attached to the head coil, which was individually adjusted for clear visibility of the screen at the back of the scanner. Participants' vision was corrected to normal using MRI-compatible sports glasses when necessary. Their responses were recorded with controllers in participants' hands (Nordic NeuroLab, Bergen, Norway). The structural scan sequence used MPRAGE T1 weighted with a repetition time TR=1.66 seconds, Echo time TE=2.54ms, flip angle = 9 degrees, FoV=256mm, and slice thickness of 0.8mm, resulting in isometric voxels of size 0.8X0.8X0.8mm. Functional scans used an EPI sequence of 37 slices, with a flip angle of 90 degrees, slice thickness of 3mm, Field of View of 192 mm, and 0.3mm distance between slices, resulting in a voxel size of 2X2X3 mm. fMRI volumes repetition time TR=2.5 seconds, echo time TE=30ms. Pre-scanning normalization and a GRAPPA factor

of 2 were used in the structural and functional sequences. Fieldmap sequences used a gradient echo sequence (GRE), which generated two magnitude images and one phase difference image.

4.8 Data Analysis

4.8.1 Measures

4.8.1.1 Behavioral Measures

Representational Dissimilarity Matrices (RDMs): Six dissimilarity matrices were calculated from the Euclidean distances between the ratings of each of the eight students over three traits (physical facial difference, personality, and social life) before and after the intervention (3 traits x 2 timepoints).

Perceived Within-Group Similarity (WG): which is the average similarity ratings of each two students that belonged to the same group for each of the three traits measured (physical facial difference, personality, and social life). WG is measured twice for each trait, once before and once after the intervention.

Perceived Between-Groups Similarity (BG): which is the average similarity of each two faces that belonged to different groups. Similar to WG, there are also six BG values (3 traits x 2 timepoints).

4.8.1.2 fMRI Measures

Discriminability Matrices; before and after-intervention LDC matrices for each searchlight (2 timepoints x 81924 searchlights)

Between-Group Discriminability (BGD): This is the average discriminability values for each two faces that belonged to different groups. There were two values for each searchlight (2 timepoints x 81924 searchlights).

Within-Group Discriminability (WGD): This is the average discriminability **value** for each two faces that belonged to the same group. There were two values for each searchlight (2 timepoints x 81924 searchlights).

4.8.2 Analysis

4.8.2.1 Behavioral Data Analysis

The average face similarity between each two faces that belong to the same group (Within-group similarity) and between each two faces that belong to different groups (intergroup similarity) will be calculated from the pairwise similarity ratings before and after

revealing the group membership of the faces. We will test hypothesis one using these variables.

4.8.2.2 fMRI Analysis

Data preparation: We started by organizing our raw MRI data into a BIDS layout (Gorgolewski et al., 2016) to facilitate compatibility with the different softwares used for the analysis.

Quality Check: We ran an MRIQC script on our data and then manually evaluated structural and functional artifacts based on the criteria mentioned in Klapwijk et al., (2019) and Provins et al., (2023). We excluded 3 participants with excessive movement (calculated as framewise displacement) and strong aliasing artifacts.

Preprocessing: fMRIPrep (version 1.0.3) was run with default settings and six degrees of freedom for alignment. Surface reconstruction was done using Freesurfer (version 6.0.1). The preprocessed data was output to the fsaverage6 template space.

Hyperalignment: The movie task data obtained in session one was used in functional alignment to estimate a common functional space. We then estimated a projection map to transform the Face Perception Task scans from the fsaverage6 space to the Hyperalignment space. Hyperaligned data were then transformed from surface-based data into volumetric data.

GLMs: Main Task GLM was created using SPM 12 (7771), data for each subject were modeled using a general linear model (GLM) using SPM12, and each session was modeled separately. For each session, we modeled trials of the six runs, each run included one regressor for each different face, in addition to 6 motion regressors, six artefact covariate regressors (CompCorr, generated by fMRIprep), and an intercept. The task-related regressors were convolved with a canonical HRF, and a high-pass filter with a cutoff period of 128 seconds was applied; no voxel masking was applied. After estimating the parameters for the GLM, the resulting beta images were projected onto the brain surface.

Defining Searchlights: Searchlights were defined on the brain surface as discs with a 10mm. After subtracting the medial brain wall from the brain surface, we were left with 40962 searchlights for each hemisphere, resulting in a total of 81924 searchlights for the whole brain. Each search light was a disc containing, on average, 180 surface nodes.

Calculating Linear Discriminant Contrast (LDCs): We calculated a matrix of linear discriminant contrasts (LDCs) in each searchlight. The LDC is a value calculated in each searchlight between each pair of the eight conditions/faces that our participants saw. This results in 28 pairs of LDC values, which we refer to as the "discriminability matrix". The discriminability matrix is a detailed representation of how strongly the neural activity elicited in each searchlight can discriminate between the eight different faces. In order to calculate LDC values, we adapted the code published by Tsantani et al., 2019. First, we extract beta values and GLM residuals in searchlight voxels. Then, the beta images are compared for each pair of conditions to evaluate how distinguishable the activity patterns are, as follows:

$$\begin{aligned} d^2_{Mahalanobis,crossvalidated}(\mathbf{b}_k, \mathbf{b}_j) &= (\mathbf{b}_j - \mathbf{b}_k)_A \Sigma_A^{-1} (\mathbf{b}_j - \mathbf{b}_k)_B^T \\ &= \mathbf{c}\mathbf{B}_A \Sigma_A^{-1} \mathbf{B}_B^T \mathbf{c}^T \\ &= LDC(\mathbf{b}_k, \mathbf{b}_j). \end{aligned}$$

Equation 1: The linear discriminant contrast equation (LDC) between any two conditions, k and j. b_k and b_j are beta values of all surface nodes of one searchlight in conditions k and b, respectively. The symbol $\sum_A^{-1}$ represents the hold-one-out cross-validation that we used.

The LDC calculation is derived from linear discriminant analysis (LDA) and aims to find a linear combination of the voxel values that best discriminates between the conditions. First, we pooled beta values from all the scanning runs collected from our 17 subjects, resulting in 101 runs (17*6 – 1 excluded run). LDC is a cross-validated measure. So, we conducted a hold-one-out cross-validation, so each LDC value is an average of 101 values. In each cross-validation run, we first calculated the Euclidean distance between the average beta values of the two conditions in the training runs (100 runs). Then, we calculated the noise covariate matrix for the two conditions. Then, the root squared distance of the training runs is divided by the noise matrix for normalization. The next step was to calculate the Euclidean distance between the beta values of the same two conditions in the test run, which was held out. Finally, the distance in the test run is multiplied by the normalized distance in the training runs.

Discriminability Matrices: The previous calculation was done for each combination of experimental contrast (each two different faces). This results in 28 discriminability values, which could be arranged as a discriminability matrix.

4.9 Hypothesis testing

4.9.1 Hypothesis 1

A linear mixed model was created to model each of the 3 dimensions (physical facial features, personality, and social life) separately. In each model we had two fixed effects factors; a) Time (pre- vs. post-intervention), and b) Rating Type (Within- vs. Between-group similarity). The model was not balanced because for each subject and each time point, we had 12 within-group but 16 between-group ratings. We also modeled the participant ID as a random effects factor. After fitting each model, we created two post-hoc contrasts to test a) whether post-intervention within-group similarity ratings were significantly higher than pre-intervention ratings, and b) whether post-intervention between-group similarity ratings were significantly lower than pre-intervention ratings.

4.9.2 Hypothesis 2

We conducted two two-sample t-tests in each of the brain's 81924 searchlights, resulting in (81924 x 2) tests. In each searchlight, we tested for the following:

Contrast 1: Which regions show social categorization after the intervention, i.e., post-intervention BGD discriminability larger than post-intervention WGD?

Contrast 2: which regions show significantly higher change in BGD than WGD from pre- to post-intervention scores?

Our tests were one-tailed because we were looking for brain regions that showed these very specific effects. The identification of voxels was implemented using image thresholding. In each of the two contrasts, only significant searchlights ($p<0.05$, right-tailed) were kept; zeros replaced other searchlights. To correct for multiple comparisons (81924, one for each searchlight), we used a non-parametric method called Threshold-Free Cluster Enhancement (TFCE). The TFCE method was designed to enhance statistical maps by considering both the signal's intensity (height) and its spatial extent, without needing to apply an arbitrary cluster-defining threshold (S. M. Smith & Nichols, 2009). We used the Matlab-TFCE toolbox version r. 269 (part of the CAT toolbox (Gaser et al., 2024)) with permutations set to 10,000.

4.9.3 Hypothesis 3

We conducted right-sided Spearman correlation tests in the whole brain (81924 searchlights) once for each post-intervention behavioral task (Physical facial features,

Personality, Social life) to test the third hypothesis. We used the same thresholding method as in hypothesis 2, where only significant searchlights in the positive direction ($p<0.05$, right-tailed) were kept. Next, we manually created a null distribution. We generated 100 datasets of null results through random permutation of the results. So, each searchlight had one result value and 100 null values for each behavioral test. Next, we conducted TFCE correction as in hypothesis 2 and used the null dataset as an input for the TFCE correction.

4.10 Results

4.10.1 Behavioral

Descriptive Statistics: Participants behaved differently when rating pairwise similarity of physical facial features in comparison to pairwise similarity of personality and social life. As shown in Figure 15, the average between-group and within-group ratings of similarity in facial features only changed slightly after the intervention. Ratings of physical facial features were as follows: between-groups (pre-intervention; mean=3.15, SD=1.5; post-intervention; mean=3.24, SD=1.5). Within-group (pre-intervention; mean= 3.6, SD= 1.7 post-intervention; mean=3.7, SD=1.6). Pairwise similarity ratings of personality and social life showed an increase in within-group similarity and a decrease in between-group similarity, depicted in Figure 16.

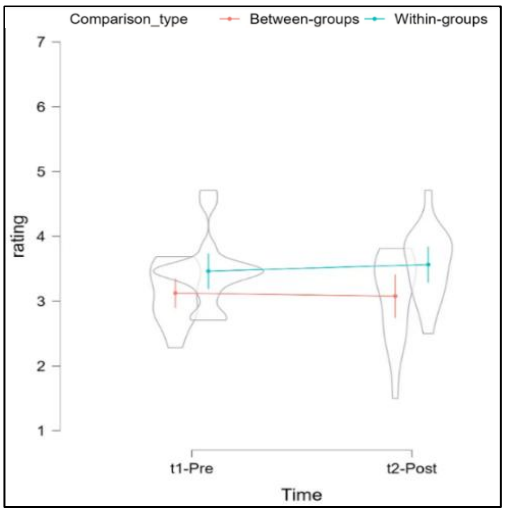


Figure 15: Pairwise similarity ratings of physical facial features before and after the intervention

Ratings of similarity in personality were as follows; between-groups (pre-intervention; mean=3.15, SD=1.5 post-intervention; mean=3.24, SD=1.5), Within-groups (pre-

intervention; mean= 3.6, SD= 1.7 post-intervention; mean=3.7, SD=1.6). Ratings of similarity in social life were as follows: between-groups (pre-intervention; mean=3.96, SD=1.5; post-intervention; mean=3.43, SD=1.35). Within-groups (pre-intervention; mean= 4.08, SD= 1.46 post-intervention; mean=4.44, SD=1.4).

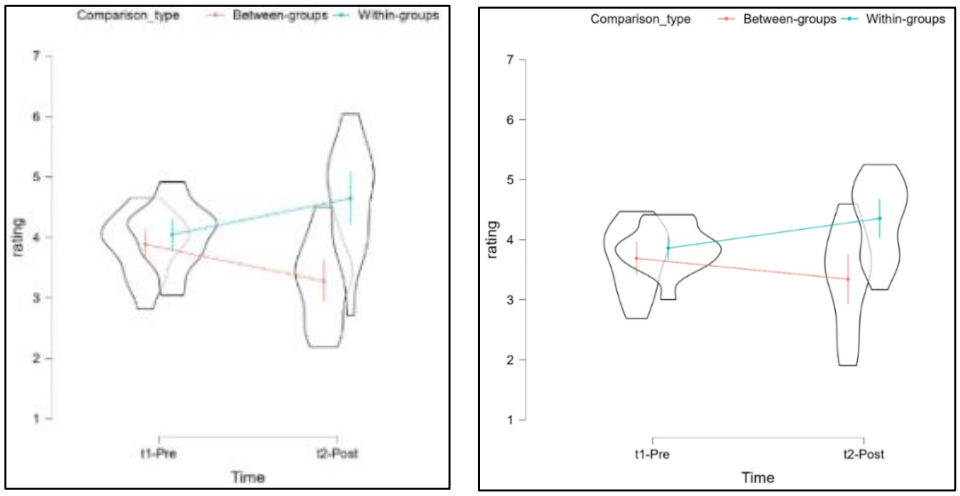


Figure 16: Pairwise similarity ratings of personality and social life before and after the intervention

Linear Mixed Models: the Linear mixed model showed a significant main effect of rating type and a significant interaction for personality and social life ratings ($p<0.001$ for both). However, no significant interaction was found between time and rating type for physical features.

Table 6: Summary of 3 ANOVA results, one for each behavioral dimension. The ANOVAs were conducted on pre- and post-intervention ratings of pairwise similarity in physical facial features, personality, and social life.

	ANOVA Summary			
	Effect	df	F	p
Physical facial features	Time	1, 16.02	0.068	0.797
	Comparison type	1, 16.01	13.903	0.002
	Time*Comparison type	1, 26.81	0.891	0.354
Personality	Effect	df	F	P
	Time	1, 19.01	1.054	0.317
	Comparison type	1, 16.06	15.755	0.001
	Time*Comparison type	1, 17.60	20.838	<.001
Social life	Effect	df	F	P
	Time	1, 16.00	0	0.992

	Comparison type	1, 16.00	25.065	< .001
	Time*Comparison type	1, 16.00	17.993	< .001

Post-hoc contrast showed a significant increase in within-group similarity ratings of personality ($p<0.001$) and social life ($p=0.003$) but not in similarity ratings of physical features ($p=0.9$). There was also a significant decrease in between-group similarity ratings of personality ($p<0.001$), social life ($p<0.001$), but not in physical features ($p = 0.9$).

4.10.2 fMRI

Hypothesis 2: A whole-brain searchlight analysis revealed that specific brain regions showed social categorization in session 3 (see Figure 17). Another set of cortical areas showed a significant decrease in between-group similarities and a significant increase in within-group similarities (see Figure 18). We created an intersection mask that revealed cortical areas with both significant effects (see Figure 19).

4.10.2.1.1.1Contrast 1: Post-intervention Social Categorization

Whole-brain searchlight analyses depicted in Figure 17 revealed a set of cortical regions that showed clear categorization between the representation of Natural science students and Arts and Humanities students. This was seen in the significantly higher between-group discriminability than within-group discriminability in post-intervention LDC matrices.

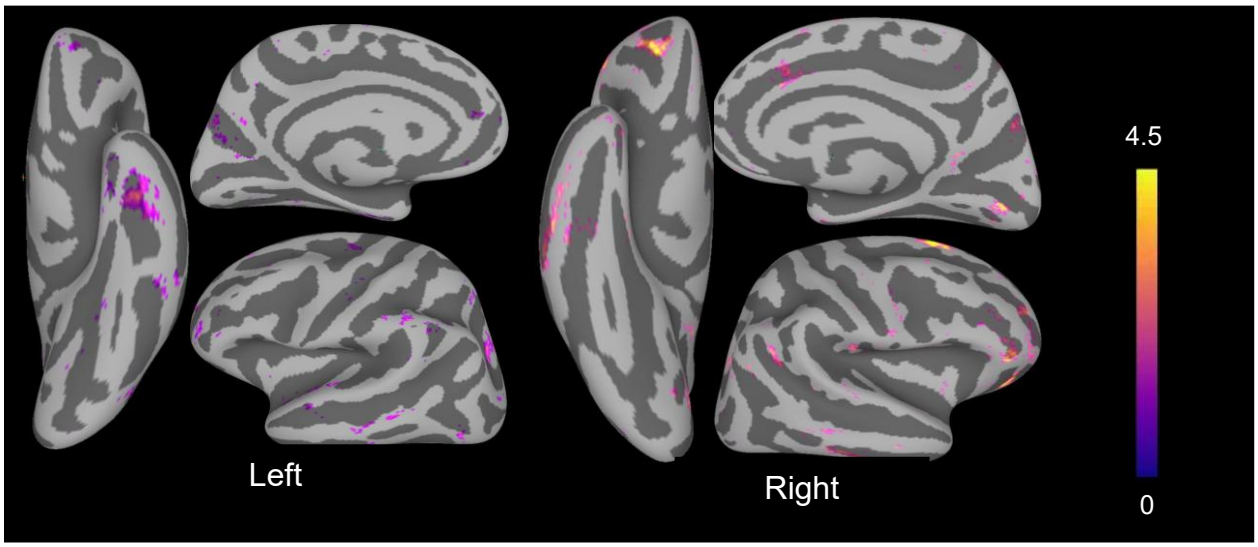


Figure 17: Searchlights which showed significantly higher between-group discriminability than within-group discriminability. The whole-brain analysis identified 907 and 630 significant searchlights in right and left hemispheres respectively. The searchlights were clustered in different brain regions all over the brain. The colorbar indicates t values.

4.10.2.1.1.2Contrast 2: Significant increase in discriminability

Whole-brain searchlight analyses results shown in Figure 18 revealed a set of cortical regions which showed a significant increase in discriminability from pre-intervention to post-intervention sessions. This could have happened through either an increase in the average between-group discriminability or a decrease in within-group discriminability or a combination of both.

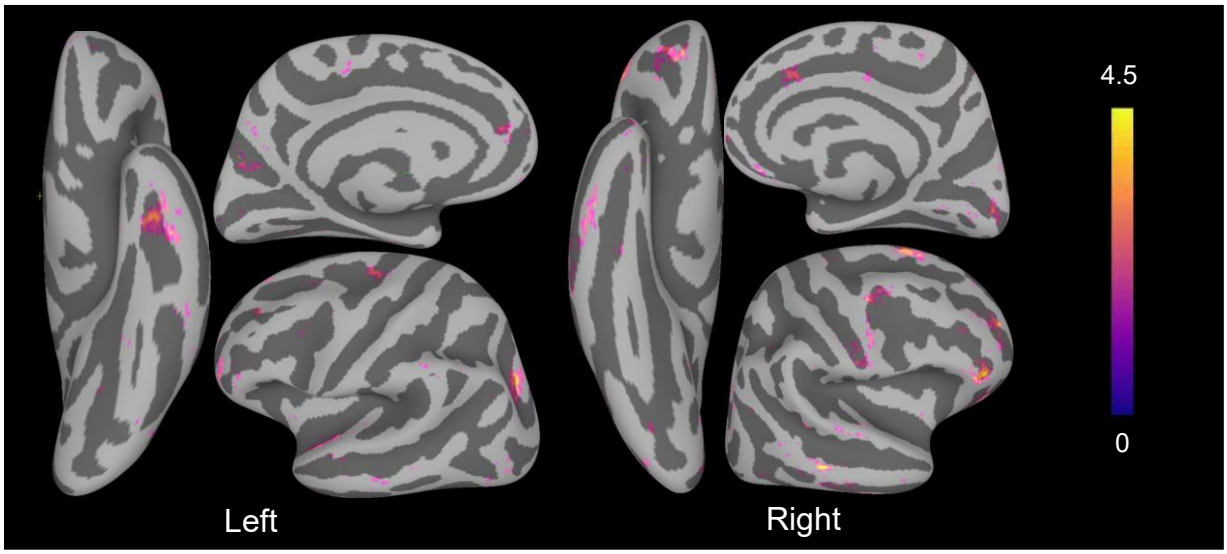


Figure 18: Searchlights showing significantly higher change in between-group discriminability than the change in within-group discriminability from pre-intervention to post-intervention scans. This whole-brain analysis revealed 770 and 694 significant searchlights in right and left hemispheres respectively. The searchlights were clustered in different brain regions all over the brain. The colorbar indicates t values.

4.10.2.1.1.3Intersection of both effects

We created a binary mask from the intersection of searchlights that were significant in the first and second effects. Then the binary mask was weighted by the first effect (post-intervention social categorization). By creating a mask that combines both effects, we are not only selecting the searchlights that showed post-intervention social categorization, but also the same searchlights changed its discriminability pattern significantly from before to after the intervention. This way, we eliminated any baseline effect from before the intervention. The binary mask was weighted by post-intervention social categorization t-value to show which searchlights that eventually showed the strongest and weakest social categorization. It's important to notice that this is a map of surface searchlights; meaning that each surface node represent the strength of the social categorization from nodes in a disc with 10 mm radius around this node. Finally, searchlights that show both effects

significantly can be considered “Neural Correlates of Social Categorization”. The intersection between the two significant effects showed activity in the Right cortex in the Superior Frontal gyrus, the inferior frontal gyrus (orbital), and the middle frontal sulcus. On the ventral side of the right cortex, Orbitofrontal and Inferior Temporal gyri. Similarly, the Left hemisphere had searchlights that correlated with “Social Categorization” in Inferior Temporal gyrus extending to the left Collateral sulcus, Middle Frontal gyrus, and Superior Occipital gyrus.

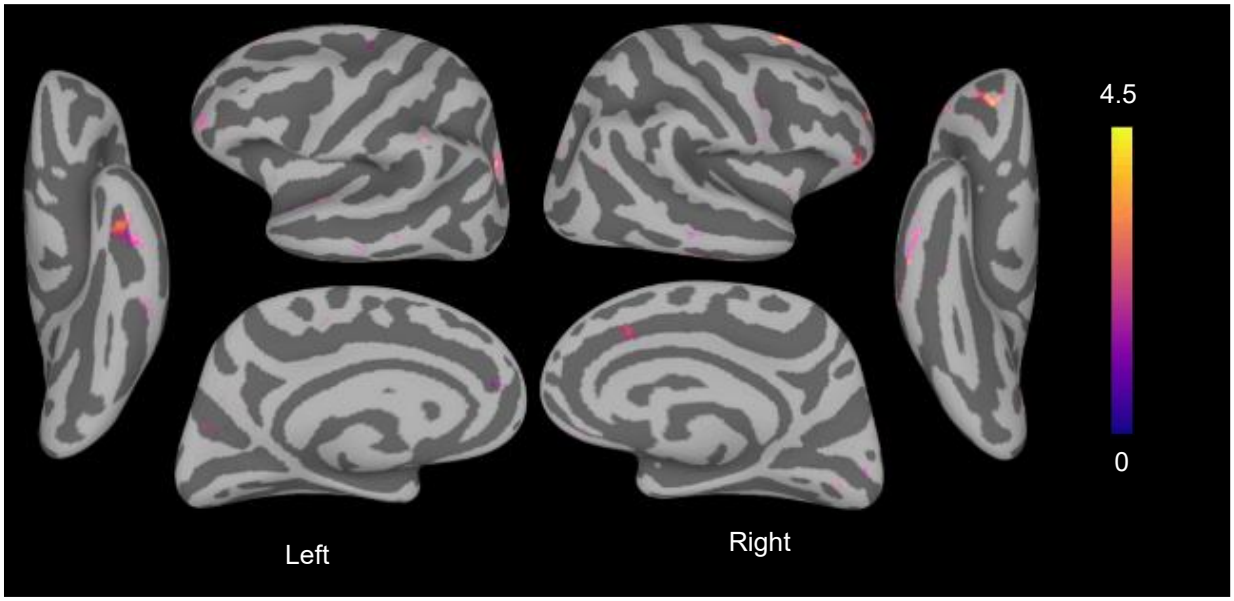


Figure 19: Searchlights showing both significantly higher increase in BGD than WGD from pre- to post-intervention scans along with significantly higher BGD than WGD after the intervention. The searchlights were identified using an intersection mask between contrast 1 and contrast 2, which revealed 395 and 302 searchlights in right and left hemispheres respectively. The searchlights were clustered mainly in frontal and temporal lobes The colorbar indicates t values.

To confirm that activation patterns in these searchlights reflect a change in discriminating between AH and NAT Students, resulting in post-intervention social categorization, we extracted the pre- and post-intervention discriminability values (the linear discriminant contrast scores) for review. Figure 19 shows the pre- and post-intervention discriminably maps in the top two regions with the highest t-values in contrast 1, i.e., the searchlights which showed the highest between-group discriminability in comparison to within-group discriminability in all the four searchlight examples shown in Figure 20.

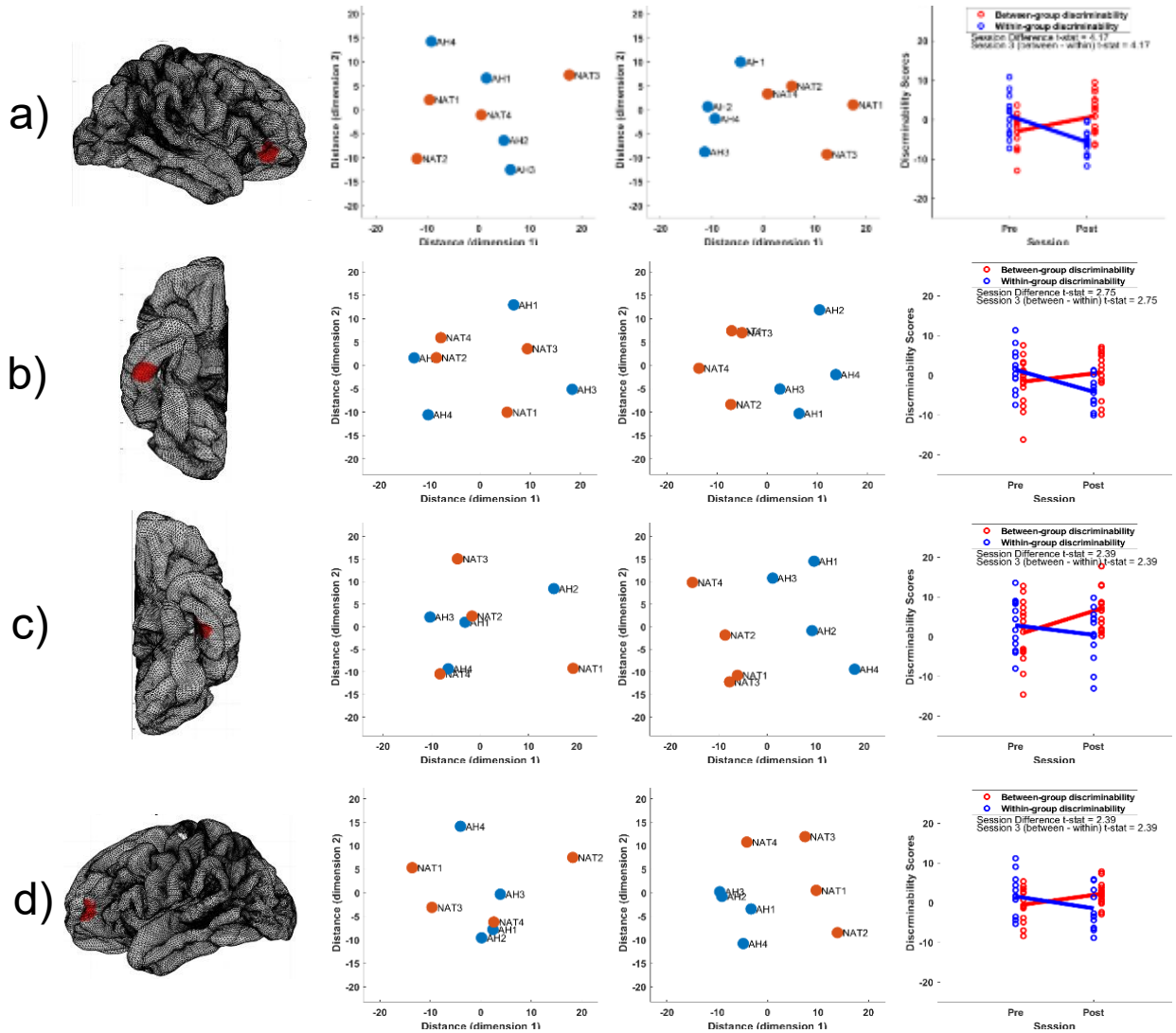


Figure 20: Discriminability change in four example searchlights (a, b, c and d) out of 697, which were identified as significant in both of the contrasts we tested. The discriminability maps in all four searchlights showed no categorization of faces into two distinct groups before the intervention (first column of MDS graphs in a, b, c and d); however, they showed clear categorization into two groups after the intervention (second column of MDS graphs). Subplots in the last column depict the underlying change in response pattern discriminability that drove the significant effect: changes from pre- to post-intervention discriminability resulted from an increase in between-group discriminability, a decrease in within-group discriminability, or both.

Hypothesis 3: We identified several cortical regions that exhibited significant correlations with the dissimilarity matrices of physical features ratings (max ρ = 0.68), personality ratings (max ρ = 0.68), and social life ratings (max ρ = 0.6).

4.10.2.1.1.4Similarity in Physical facial features

Table 7: Correlation between brain activation patterns and ratings of physical facial features

Physical facial features, Right hemisphere	**Correlation**
Dorsomedial posterior superior frontal gyrus	0.68
Middle Superior Frontal Gyrus and Sulcus, extending ventrally to middle frontal gyrus and sulcus, inferior frontal sulcus and opercular part of IFG	0.66
Inferior Frontal Triangular Gyrus and Sulcus	0.63
Posterior Lateral Fissure, planum temporal	0.57
Posterior Superior Temporal Sulcus	0.55
Marginal Sulcus of the Cingulate	0.54
Angular Gyrus extending to middle occipital	0.53
Middle Occipital and Lunatus Sulcus	0.53
Gyrus Rectus and Orbital gyrus	0.5
Superior Paracentral Gyrus	0.5
Anteromedial part of superior frontal gyrus	0.44
Physical facial features, Left hemisphere	
Temporal Pole	0.63
Insular Gyrus and Central Sulcus of the Insula	0.61
Middle part of Superior Temporal Sulcus and gyrus	0.6
Posterior Lateral Fissure	0.54
Anterior parahippocampal gyrus	0.53
Occiptio-temporal lingual gyrus	0.5

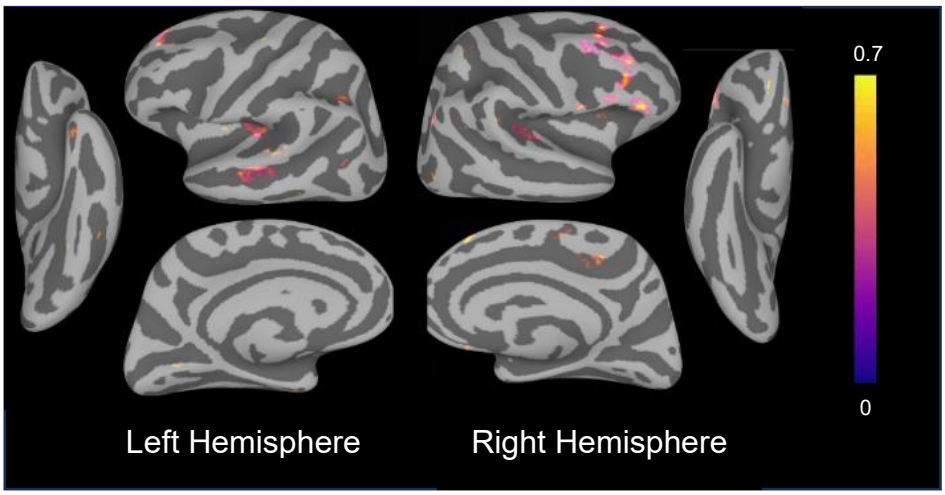


Figure 21: Results of whole-brain searchlight correlation between discriminability matrices in each searchlight and pairwise similarity ratings of physical facial features among the eight students shown in the scanner. There were 774 and 578 significant searchlights in the right and left hemispheres respectively. The colorbar indicates Spearman's correlation values ρ.

4.10.2.1.1.5Similarity in Personality

Table 8: Correlation between brain activation patterns and ratings of personality similarity

Personality, Right hemisphere	**Correlation**
Insular Gyrus and Central Sulcus of the Insula	0.68
Superior Frontal Gyrus and Sulcus	0.59
Superior Frontal Sulcus	0.56
Planum Temporale and posterior sylvian fissure	0.54
Inferior Temporal Sulcus	0.5
Anterior Superior Temporal Gyrus	0.48
Intraparietal and Transverse Parietal Sulcus	0.46
Personality, Left hemisphere	
Posterior Superior Temporal Gyrus.	0.62
Anterior Superior Temporal Sulcus	0.58
Gyrus Rectus and Sulcus	0.58
Anterior Occipital Sulcus	0.56
Posterior Insular Gyrus	0.54
Medial posterior Superior Temporal gyrus and Sulcus	0.42

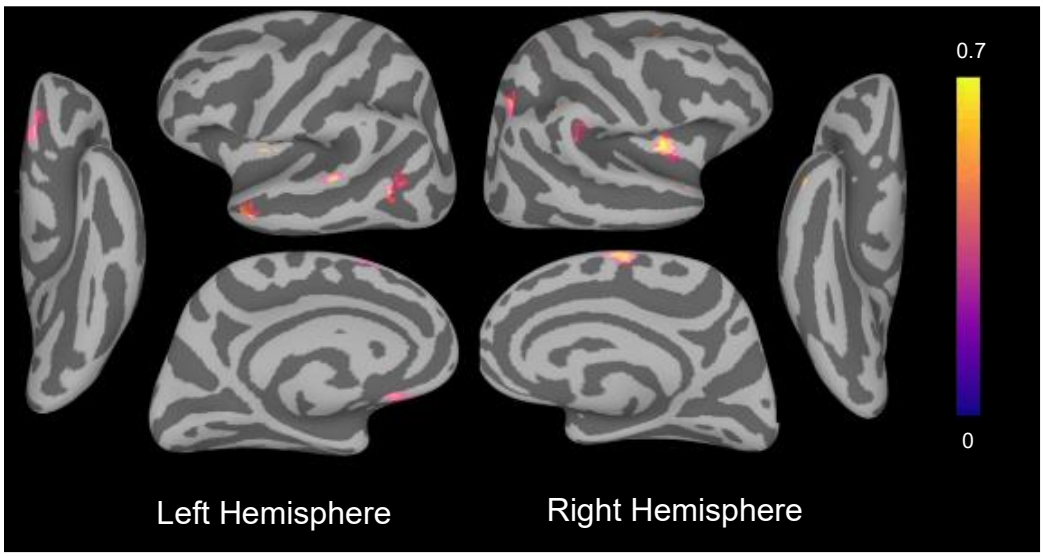


Figure 22: Results of whole-brain searchlight correlation between discriminability matrices in each searchlight and pairwise similarity ratings of personality among the eight students shown in the scanner. There were 304 and 505 significant searchlights in the right and left hemispheres, respectively. The colorbar indicates Spearman's correlation values ρ.

4.10.2.1.1.6Similarity in Social life

Table 9: Correlation between brain activation patterns and ratings of social life similarity

Social life, Right hemisphere	**Correlation**
Insular Gyrus and Central Sulcus of the Insula	0.59
Middle Occipital Gyrus and Sulcus	0.59
Planum Temporale (Superior Temporal Gyrus and Sulcus) & Inferior Parietal (Supramarginal) Gyrus and Sulcus	0.54
Middle to Posterior Cingulate Sulcus	0.52
Anterior Superior Temporal Gyrus and Lateral Sulcus	0.51
Social life Left hemisphere	
Posterior Superior Temporal Gyrus and Posterior Lateral Sulcus	0.6
Precentral Gyrus and Sulcus	0.54
Intraparietal and Transverse Parietal Sulcus	0.53
Superior Part of the Precentral Sulcus	0.52
Anterior Superior Temporal Sulcus	0.488
Inferior Parietal (Supramarginal) Gyrus and Sulcus	0.47
Posterior Insular gyrus	0.47
Postcentral Sulcus	0.45

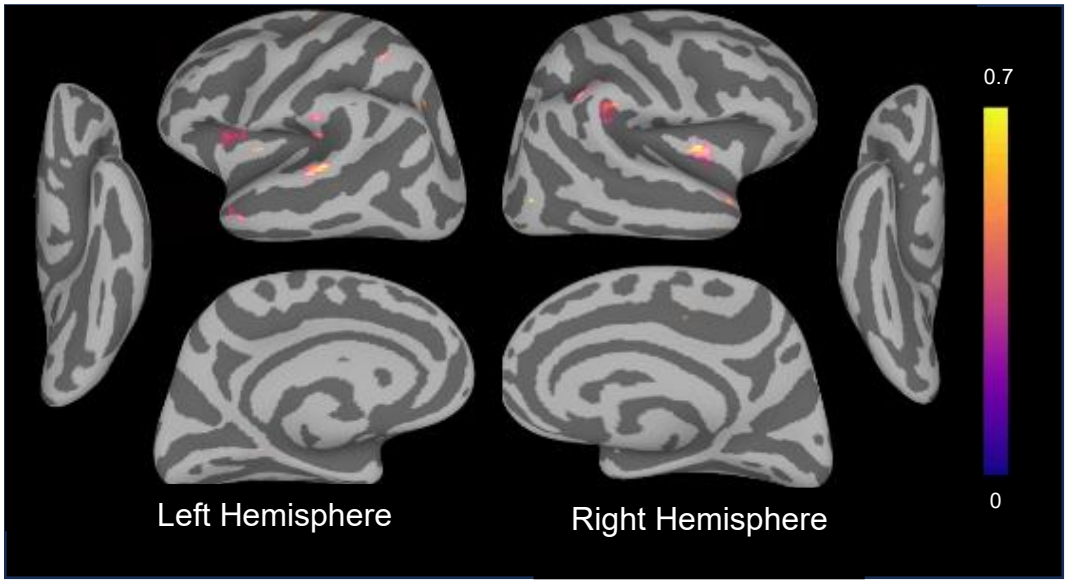


Figure 23: Results of whole-brain searchlight correlation between discriminability matrices in each searchlight and pairwise similarity ratings of social life among the eight students shown in the scanner. There were 288 and 355 significant searchlights in the right and left hemispheres respectively. The colorbar indicates Spearman's correlation values ρ.

4.11 Discussion

This study investigated how revealing social group membership influences perceived similarity in personality, social life, and facial features, as well as how these changes are reflected in neural representations. The findings provide robust support for the hypothesis that social categorization alters both behavioral judgments and brain activity.

4.11.1 Hypothesis 1: Behavioral Changes Due to Social Categorization

Our results revealed a clear pattern: post-intervention, participants rated within-group pairs as more similar and between-group pairs as less similar regarding personality and social life, but not in physical facial features. This supports the idea that social categorization modulates higher-order social judgments rather than basic perceptual attributes. The lack of significant change in physical facial similarity ratings suggests that visual feature-based judgments are relatively stable and less susceptible to group-based biases, consistent with earlier findings on the separability of perceptual and social categorization mechanisms (Van Bavel et al., 2008). The significant interaction between time and comparison type for personality and social life ratings suggests that the mere knowledge of group membership was sufficient to shift perceived interpersonal similarity along social dimensions.

4.11.2 Hypothesis 2: Neural Correlates of Social Categorization

Hypothesis 2 proposed that making social group membership visible would change neural similarity patterns in brain areas related to how we see and judge other people. The multivariate pattern results support this idea, showing that patterns of neural activity changed in several distinct brain areas. These changes likely reflect the influence of different mental processes, such as visual perception, emotional evaluation, and cognitive control.

Perceptual and Semantic Processes: The Left Superior Occipital Gyrus (SOG*)* is usually involved in processing spatial aspects of visual information and in visual memory. Its activity in this study may reflect fine spatial processing of facial features or gaze direction cues that may become more important after social labels are applied. A low-level visual area like the SOG shows these effects, suggesting that high-level social knowledge can influence even the early stages of perception.

The Inferior Temporal Gyrus (ITG) is increasingly recognized for its dual role in processing visually and semantically rich stimuli, including social cues. Although the region is not considered part of the canonical face-processing network (e.g., FFA or OFA), several studies highlighted the ITG's involvement in semantic and social cognition. Binder et al. (2009) conducted a meta-analysis revealing consistent FG/ITG activation during semantic tasks, supporting its function in integrating visual and conceptual information. Similarly, Visser et al. (2010) found bilateral anterior temporal lobe activation, including the ITG, during time-constrained semantic categorization. Shkurko et al. (2013) reported that this region, while influenced by top-down processes, responds to outgroup stimuli and to non-facial, socially meaningful cues (Hein et al., 2010), indicating a role in encoding identity-relevant social information. Furthermore, Jonas et al., (2016) documented face-selective responses in the anterior ITG and MTG, suggesting these areas contribute to facial identity representation by linking perceptual and semantic data. Overall, these findings argue for the ITG's broader function in integrating perceptual and conceptual aspects of socially salient stimuli, especially in person perception and categorization contexts.

Cognitive Functions and working memory activation: Activity in the Right Middle Frontal Sulcus and Left Middle Frontal Gyrus (MFG), both of which are considered parts of the Dorsolateral Prefrontal Cortex (DLPFC), points to increased cognitive control. These areas are often involved in explicit or conscious behaviors and thoughts. Processes like suppressing automatic reactions or updating old beliefs towards a social group would likely engage the DLPFC (Forbes & Grafman, 2013).

Although the primary goal of our fMRI paradigm was to study face perception, the nature of our task – the "Face Perception Task"- inherently involved working memory (WM) processes, such as encoding, maintenance, and comparison of face stimuli. Therefore, one would expect consistent recruitment of core WM-related regions—such as the middle frontal gyrus (MFG) and right superior frontal gyrus (SFG)—across all trials, irrespective of the social attributes of the faces being evaluated. In line with this, Yaple et al. (2019) reported that both right MFG and right SFG are frequently activated during working memory tasks.

Therefore, the modulation in activation patterns within these regions following the social categorization intervention is theoretically unexpected under a purely domain-general WM

account. Notably, however, regions associated with executive control (rMFG, rSFG) and attentional switching (right inferior frontal gyrus, RIFG) exhibited differential activation patterns depending on the group membership of the perceived face. This finding suggests that working memory processes may be susceptible to top-down modulation by the social category or attributes of the stimulus. These effects appeared even though the physical appearance of the faces remained unchanged. This supports the idea that socially salient cues—such as group membership— can reshape how we see and remember others.

4.11.3 Hypothesis 3: Correlation Between Neural Discriminability and Behaviorally Reported Dissimilarity Maps

In Hypothesis 3, we focused more on the post-intervention behaviorally reported perceived dissimilarity between the eight students in terms of physical facial features, personality, and social life. In this investigation, we looked for brain searchlights that showed post-intervention dissimilarity patterns of fMRI activity that were positively correlated with the behaviorally reported patterns. Therefore, regions that showed a significant positive correlation with the behaviorally reported dissimilarity maps were able to track and represent the perceived group topology over different behavioral dimensions.

We searched for regions whose neural discriminability patterns correlated with behaviorally reported perceived dissimilarity maps across all three dimensions. No such brain regions were identified, suggesting that no neural system uniformly tracked interpersonal similarity across both semantic and perceptual domains. However, we found significant correlations between neural and behavioral dissimilarity patterns of personality and social life in the right insula, posterior segment of the right lateral fissure, left posterior superior temporal gyrus (STG), and left anterior superior temporal sulcus (STS). These areas did not show similar correlations with dissimilarity ratings in physical facial features, supporting a dissociation between neural representations of perceptual and semantic trait information.

The involvement of these regions aligns well with existing literature on the neural basis of social cognition. The insula has been implicated in emotional awareness and interoceptive processing, emotional responses in social interactions (Craig, 2009; Uddin et al., 2017; Boucher et al, 2015). The anterior insula has also been linked to social affect, including empathy, emotional salience, and norm violation detection (Gu et al., 2013;

Lamm & Singer, 2010). Anterior STS plays a key role in encoding social salience and has been shown to track socially meaningful distinctions even in the absence of overt task demands. The anterior STS is implicated in social tasks like social concepts (Zahn et al., 2007), semantic judgments about people (Mitchell et al., 2002), and voice-identity (Perrodin et al., 2015). The region in the left STG seems, however, to be posterior to the areas involved in semantic processing, as it is rather known for its role in speech processing (Ozker et al., 2024).

The current results extend this literature by showing that after social group membership is revealed, these areas appear to encode a similarity structure that mirrors behavioral impressions, even though participants were not explicitly asked to make semantic evaluations during scanning. This suggests that the brain constructs and maintains socially meaningful representational spaces that align with observers' perceptions of others' traits, particularly along dimensions shaped by group membership. Although these regions discriminated between the faces of the students in a way that correlated positively with behaviorally reported perceived dissimilarity in social life and personality (semantic judgments) and not in physical facial feature (perceptual judgments), we cannot rule out that the correlation in these common regions could be because the behaviorally reported dissimilarity pattern in post-intervention ratings showed social categorization only for semantic judgments but not for perceptual judgments, which means that the pattern of neural representations might reflect ingroup/outgroup distinctions, rather than a finer-grained encoding of trait dissimilarity.

4.11.4 Theoretical Implications

These findings contribute to a growing body of literature demonstrating the flexibility of social cognition. Using a controlled intervention (revealing group membership of newly encountered students), we managed to causally link cognitive group labeling with both perceptual judgment and neural representation changes. Moreover, the separation of traits into perceptual (physical facial features) and semantic (personality and social life) enables a better understanding of how different levels of social information are processed. Notably, the results support a dynamic model where social information shapes the perception of persons.

4.11.5 Limitations and Future Directions

Despite these insights, several limitations must be acknowledged. First, the sample size, while within norms for neuroimaging studies, limits generalizability. Second, the use of university-based group labels (Natural Sciences vs. Arts and Humanities) may not evoke strong real-world social identities, potentially underestimating effects. Future studies could incorporate more emotionally salient or identity-relevant groupings to test the boundaries of these effects. Additionally, while the searchlight analysis identified cortical regions involved in group-based categorization, further work is needed to map and understand the specific functional roles of these areas from a network perspective. And to integrate this research with previous literature.

4.11.6 Conclusion

This study advances our understanding of how group membership information reshapes both subjective social perception and the neural representation of others. By linking behavioral and neural data through a robust experimental design, we provide compelling evidence for the flexibility of social categorization mechanisms and their impact on interpersonal cognition.

5 Discussion and Conclusion

Throughout this thesis, we focused on studying the perception of social groups from an individual's point of view. We approached the topic from both psychological and neurological perspectives. We measured how students were perceived before and after being assigned to a social group. We also measured the outgroup homogeneity effect, the perceived similarity between individuals over a perceptual trait (physical facial similarity) and more semantic traits (personality and social life). Most importantly, we compared perceptual ratings of students resulting from group-level rating tasks to ratings resulting from individual-level rating tasks.

In the first study, we started by replicating an experiment by Pickett & Brewer (Pickett & Brewer, 2001), hoping to replicate the outgroup homogeneity effect (Ostrom & Sedikides, 1992) and the influence of social identity threat on perceived ingroup and outgroup homogeneity. Unfortunately, we got mixed results from the first experiment, where we managed to replicate the outgroup homogeneity effect using group-level tasks, but we couldn’t replicate the effect of social identity threat on the perceived outgroup homogeneity. Most importantly, we discovered a discrepancy between individual-level and group-level tasks of social group perception. This discrepancy motivated us to focus on individual-level tasks, which were more appropriate to conduct when combining behavioral measurement with functional MRI scanning. Using the pairwise similarity task developed in Study 2, we managed to behaviorally measure social categorization caused by revealing the group membership of newly encountered students.

Following this finding, we progressed to add an fMRI task before and after revealing the group membership of the students in our third study. We managed again to find the same social categorization effect as in Study 2. In addition, we found a set of brain regions in the bilateral ITG and DLFPC, and the right IFG and OFC, which changed their neural representations to the students’ faces after revealing the group membership of the students. Our intervention caused these regions to either a) show more similar representations between the faces belonging to the same group, or b) decrease similarity between representations of faces belonging to different groups, or c) both effects. This resulted eventually in these regions showing social categorization, where the ability to discriminate between faces belonging to different groups is significantly higher compared to faces belonging to the same group. In addition, we also identified a set of regions that

exhibited a post-interventional discriminability pattern between faces, which significantly correlated with the behaviorally reported perceived dissimilarity in personality and social life. These regions included the right insula, left anterior STS, and left posterior STG.

The importance of these studies lies not only in their results but also in the opportunities they present. We claim that studying group perception on the individual level, using the pairwise similarity concept, provides an opportunity for an indirect and replicable mapping of the perceived social group structure over any trait in question. Another advantage is the compatibility of the pairwise similarity concept with representational similarity analysis (RSA), which allows for multimodal integration and correlation of results.

6 Abstract

This thesis investigates the cognitive and neural mechanisms underpinning the perception of group homogeneity and the similarity between individuals belonging to the same or different social groups. Across three empirical studies, it explores how social identity and categorization influence both behavioral judgments and brain activation patterns associated with perceived similarity.

Study 1 replicates and extends foundational work on Optimal Distinctiveness Theory, testing how identity threat modulates perceived similarity and stereotypicality in in-group and out-group evaluations. Despite using both group-based and individual-based tasks, the study found only partial support for the predicted increase in perceived out-group homogeneity, while revealing nuanced differences based on measurement type and participant identification strength. Study 2 introduces a novel pairwise similarity rating paradigm to quantify fine-grained behavioral ratings of perceived similarity between individuals belonging to two social groups. The study establishes this method's effectiveness in measuring social categorization of individuals and demonstrates how information about group membeship of individuals influence perceived similarity judgments between these individuals, even on traits which are indirectly related to the group membeship. Study 3 combines behavioral similarity measures with fMRI to identify brain regions encoding social categorization. Using representational similarity analysis (RSA), it reveals that the right orbitofrontal cortex, bilateral inferior temporal gyri,and the right inferior frontal and bilateral middle frontal gyri are sensitive to social group and reflect changes in perceived similarity induced by social categorization.

Together, these studies advance our understanding of how the human brain dynamically constructs and updates social similarity representations. The findings highlight the interplay between identity processes, cognitive evaluations, and neural coding, contributing to theories of social cognition and the neural basis of categorization.

7 List of Figures

8 List of Tables

9 References

Abrams, D., & Hogg, M. A. (Eds.). (1990). *Social identity theory: Constructive and critical advances* (pp. viii, 297). Springer-Verlag Publishing.

Aguirre, G. K., Mattar, M. G., & Magis-Weinberg, L. (2011). De Bruijn cycles for neural decoding. *NeuroImage*, *56*(3), 1293–1300. https://doi.org/10/bf7jq2

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience*, *15*(10), 670–682. https://doi.org/10.1038/nrn3800

Andreella, A., Finos, L., & Lindquist, M. A. (2022). Enhanced hyperalignment via spatial prior information. *Human Brain Mapping*, *44*(4), 1725–1740. https://doi.org/10.1002/hbm.26170

Anzellotti, S., Fairhall, S. L., & Caramazza, A. (2014). Decoding Representations of Face Identity That are Tolerant to Rotation. *Cerebral Cortex*, *24*(8), 1988–1995. https://doi.org/10.1093/cercor/bht046

Bahrami, M., Laurienti, P. J., Shappell, H. M., & Simpson, S. L. (2023). Brain Network Analysis: A Review on Multivariate Analytical Methods. *Brain Connectivity*, *13*(2), 64–79. https://doi.org/10.1089/brain.2022.0007

Baseler, H. A., Harris, R. J., Young, A. W., & Andrews, T. J. (2014). Neural responses to expression and gaze in the posterior superior temporal sulcus interact with facial identity. *Cerebral Cortex (New York, N.Y.: 1991)*, *24*(3), 737–744. https://doi.org/10.1093/cercor/bhs360

Bin Meshar, M., Stolier, R. M., & Freeman, J. B. (2021). Facial Stereotyping Drives Judgments of Perceptually Ambiguous Social Groups. *Social Psychological and Personality Science*, 19485506211062285. https://doi.org/10.1177/19485506211062285

Boldry, J. G., Gaertner, L., & Quinn, J. (2007). Measuring the measures: A meta-analytic investigation of the measures of outgroup homogeneity. *Group Processes and Intergroup Relations*, *10*(2), 157–178. https://doi.org/10.1177/1368430207075153

Bressler, S. L., & Menon, V. (2010). Large-scale brain networks in cognition: Emerging methods and principles. *Trends in Cognitive Sciences*, *14*(6), 277–290. https://doi.org/10/fcjtqf

Brewer, M. B. (1991). The Social Self: On Being the Same and Different at the Same Time. *Personality and Social Psychology Bulletin*, *17*(5), 475–482. https://doi.org/10.1177/0146167291175001

Brewer, M. B., & Pickett, C. L. (1999). Distinctiveness motives as a source of the social self. In *The psychology of the social self* (pp. 71–87). Lawrence Erlbaum Associates Publishers.

Bruneau, E. (2018). *Denying Humanity: The Distinct Neural Correlates of Blatant Dehumanization Denying Humanity: The Distinct Neural Correlates of Blatant Dehumanization*. *June*. https://doi.org/10.1037/xge0000417

Cao, Y., Contreras-Huerta, L. S., McFadyen, J., & Cunnington, R. (2015). Racial bias in neural response to others' pain is reduced with other-race contact. *Cortex*, *70*, 68–78. https://doi.org/10.1016/j.cortex.2015.02.010

Carlson, S. (1998). Distribution of cortical activation during visuospatial n-back tasks as revealed by functional magnetic resonance imaging. *Cerebral Cortex*, *8*(8), 743–752. https://doi.org/10.1093/cercor/8.8.743

Carrington, S. J., & Bailey, A. J. (2009). Are there theory of mind regions in the brain? A review of the neuroimaging literature. *Human Brain Mapping*, *30*(8), 2313–2335. https://doi.org/10.1002/hbm.20671

Castello, M. V. di O., Haxby, J. V., & Gobbini, M. I. (2021). Shared neural codes for visual and semantic information about familiar faces in a common representational space. *Proceedings of the National Academy of Sciences*, *118*(45). https://doi.org/10/gncxv2

Chavez, R. S., & Heatherton, T. F. (2015). Representational Similarity of Social and Valence Information in the Medial pFC. *Journal of Cognitive Neuroscience*, *27*(1), 73–82. https://doi.org/10.1162/jocn_a_00697

Collins, J. A., Koski, J. E., & Olson, I. R. (2016). More Than Meets the Eye: The Merging of Perceptual and Conceptual Knowledge in the Anterior Temporal Face Area. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00189

Collins, J. A., & Olson, I. R. (2014). Beyond the FFA: The role of the ventral anterior temporal lobes in face processing. *Neuropsychologia*, *61*, 65–79. https://doi.org/10.1016/j.neuropsychologia.2014.06.005

Contreras-Huerta, L. S., Hielscher, E., Sherwell, C. S., Rens, N., & Cunnington, R. (2014). Intergroup relationships do not reduce racial bias in empathic neural responses to pain. *Neuropsychologia*, *64*, 263–270. https://doi.org/10/f6sbgm

Crisp, R. J., Turner, R. N., & Hewstone, M. (2010). Common ingroups and complex identities: Routes to reducing bias in multiple category contexts. *Group Dynamics: Theory, Research, and Practice*, *14*(1), 32–46. https://doi.org/10.1037/a0017303

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable Neural Components in the Processing of Black and White Faces. *Psychological Science*, *15*(12), 806–813. https://doi.org/10/d5sh6b

Davis, T., LaRocque, K. F., Mumford, J., Norman, K. A., Wagner, A. D., & Poldrack, R. A. (2014). What Do Differences Between Multi-voxel and Univariate Analysis Mean? How Subject-, Voxel-, and Trial-level Variance Impact fMRI Analysis. *NeuroImage*, *97*, 271–283. https://doi.org/10.1016/j.neuroimage.2014.04.037

Dove, A., Pollmann, S., Schubert, T., Wiggins, C. J., & Yves von Cramon, D. (2000). Prefrontal cortex activation in task switching: An event-related fMRI study. *Cognitive Brain Research*, *9*(1), 103–109. https://doi.org/10.1016/S0926-6410(99)00029-4

Dovidio, J. F., Gaertner, S. L., Hodson, G., Riek, B. M., Johnson, K. M., & Houlette, M. (2006). Recategorization and crossed categorization: The implications of group

salience and representations for reducing bias. In *Multiple social categorization: Processes, models and applications* (pp. 65–89). Psychology Press.

Farmer, H., Hewstone, M., Spiegler, O., Morse, H., Saifullah, A., Pan, X., Fell, B., Charlesford, J., & Terbeck, S. (2020). Positive intergroup contact modulates fusiform gyrus activity to black and white faces. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-59633-9

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

Feng, L., Liu, J., Wang, Z., Li, J., Li, L., Ge, L., Tian, J., & Lee, K. (2011). The other face of the other-race effect: An fMRI investigation of the other-race face categorization advantage. *Neuropsychologia*, *49*(13), 3739–3749. https://doi.org/10/czb78s

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Forbes, C. E., & Grafman, J. (2013). Social neuroscience: The second phase. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00020

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

Gaser, C., Dahnke, R., Thompson, P. M., Kurth, F., Luders, E., & the Alzheimer's Disease Neuroimaging Initiative. (2024). CAT: A computational anatomy toolbox for the analysis of structural MRI data. *GigaScience*, *13*, giae049. https://doi.org/10.1093/gigascience/giae049

Gilbert, G. M. (1951). Stereotype persistance and change among college students. *Journal of Abnormal and Social Psychology*, *46*(2), 245–254. https://doi.org/10.1037/h0053696

Gilbert, S. J., Swencionis, J. K., & Amodio, D. M. (2012). Evaluative vs. trait representation in intergroup social judgments: Distinct roles of anterior temporal lobe and prefrontal cortex. *Neuropsychologia*, *50*(14), 3600–3611. https://doi.org/10.1016/j.neuropsychologia.2012.09.002

Goesaert, E., & Beeck, H. P. O. de. (2013). Representations of Facial Identity Information in the Ventral Visual Stream Investigated with Multivoxel Pattern Analyses. *Journal of Neuroscience*, *33*(19), 8549–8558. https://doi.org/10/f4vzwz

Golby, A. J., Gabrieli, J. D. E., Chiao, J. Y., & Eberhardt, J. L. (2001). Differential responses in the fusiform region to same-race and other-race faces. *Nature Neuroscience*, *4*(8), 845–850. https://doi.org/10.1038/90565

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., … Poldrack, R. A. (2016). The brain imaging data structure, a

format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*, 160044. https://doi.org/10.1038/sdata.2016.44

Gower, J. C. (1966). *Some distance properties of latent root and vector methods used in multivariate analysis | Biometrika | Oxford Academic*. https://academic.oup.com/biomet/article-abstract/53/3-4/325/246598

Guntupalli, J. S., Hanke, M., Halchenko, Y. O., Connolly, A. C., Ramadge, P. J., & Haxby, J. V. (2016). A Model of Representational Spaces in Human Cortex. *Cerebral Cortex*, *26*(6), 2919–2934. https://doi.org/10.1093/cercor/bhw068

Hampshire, A., & Owen, A. M. (2006). Fractionating Attentional Control Using Event-Related fMRI. *Cerebral Cortex*, *16*(12), 1679–1689. https://doi.org/10.1093/cercor/bhj116

Han, S. (2018). Neurocognitive Basis of Racial Ingroup Bias in Empathy. *Trends in Cognitive Sciences*, *22*(5), 400–421. https://doi.org/10/gdfdg2

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the Lowest of the Low: Neuroimaging Responses to Extreme Out-Groups. *Psychological Science*, *17*(10), 847–853. https://doi.org/10.1111/j.1467-9280.2006.01793.x

Hart, A. J., Whalen, P. J., Shin, L. M., McInerney, S. C., Fischer, H., & Rauch, S. L. (2000). Differential response in the human amygdala to racial outgroup vs ingroup face stimuli. *Neuroreport*, *11*(11), 2351–2355. https://doi.org/10.1097/00001756-200008030-00004

Haslam, N., & Stratemeyer, M. (2016). Recent research on dehumanization. *Current Opinion in Psychology*, *11*, 25–29. https://doi.org/10.1016/j.copsyc.2016.03.009

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), Article 5539. https://doi.org/10.1126/science.1063736

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B., Gobbini, M. I., Hanke, M., & Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, *72*(2), 404–416. https://doi.org/10.1016/j.neuron.2011.08.026

Haxby, J. V., James V. Haxby, Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2020). Hyperalignment: Modeling Shared Information Encoded in Idiosyncratic Cortical Topographies. *eLife*, *9*. https://doi.org/10.7554/elife.56601

Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural Responses to Ingroup and Outgroup Members' Suffering Predict Individual Differences in Costly Helping. *Neuron*, *68*(1), 149–160. https://doi.org/10/b6w6z4

Holyoak, K. J., & Morrison, R. G. (2012). The Oxford Handbook of Thinking and Reasoning. In *The Oxford Handbook of Thinking and Reasoning*. https://doi.org/10.1093/oxfordhb/9780199734689.001.0001

Jack, R. E., & Schyns, P. G. (2015). The Human Face as a Dynamic Tool for Social Communication. *Current Biology*, *25*(14), R621–R634. https://doi.org/10.1016/j.cub.2015.05.052

Jonas, J., Jacques, C., Liu-Shuang, J., Brissart, H., Colnat-Coulbois, S., Maillard, L., & Rossion, B. (2016). A face-selective ventral occipito-temporal map of the human brain with intracerebral potentials. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(28), E4088-4097. https://doi.org/10.1073/pnas.1522033113

Kahnt, T., Heinzle, J., Park, S. Q., & Haynes, J.-D. J.-D. (2010). The neural code of reward anticipation in human orbitofrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(13), 6010–6015. https://doi.org/10.1073/pnas.0912838107

Karasawa, M., Karasawa, K., & Hirose, Y. (2004). Homogeneity perception as a reaction to identity threat: Effects of status difference in a simulated society game. *European Journal of Social Psychology*, *34*(5), 613–625. https://doi.org/10.1002/ejsp.219

Karlins, M., Coffman, T. L., & Walters, G. (1969). On the fading of social stereotypes: Studies in three generations of college students. *Journal of Personality and Social Psychology*, *13*(1), 1–16. https://doi.org/10.1037/h0027994

Katsumi, Y., & Dolcos, S. (2018). Neural Correlates of Racial Ingroup Bias in Observing Computer-Animated Social Encounters. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00632

Katz, D., & Braly, K. (1933). Racial stereotypes of one hundred college students. *Journal of Abnormal and Social Psychology*, *28*(3), 280–290. https://doi.org/10.1037/h0074049

Kersbergen, I., & Robinson, E. (2019). Blatant Dehumanization of People with Obesity. *Obesity*, *27*(6), 1005–1012. https://doi.org/10.1002/oby.22460

Kim, H., Kim, G. Y., & Lee, S.-H. (2019). Effects of individuation and categorization on face representations in the visual cortex. *Neuroscience Letters*, *708*, 134344. https://doi.org/10.1016/j.neulet.2019.134344

Klapwijk, E. T., Van de Kamp, F., Meulen, M. van der, Peters, S., & Wierenga, L. M. (2019). Qoala-T: A supervised-learning tool for quality control of FreeSurfer segmented MRI data. *NeuroImage*, *189*, 116–129. https://doi.org/10.1016/j.neuroimage.2019.01.014

Klein-Flügge, M. C., Barron, H. C., Brodersen, K. H., Dolan, R. J., & Behrens, T. E. J. (2013). Segregated Encoding of Reward–Identity and Stimulus–Reward Associations in Human Orbitofrontal Cortex. *The Journal of Neuroscience*, *33*(7), 3202–3211. https://doi.org/10.1523/JNEUROSCI.2532-12.2013

Koban, L., Pichon, S., & Vuilleumier, P. (2014). Responses of medial and ventrolateral prefrontal cortex to interpersonal conflict for resources. *Social Cognitive and Affective Neuroscience*, *9*(5), 561–569. https://doi.org/10.1093/scan/nst020

Krautheim, J. T., Straube, B., Dannlowski, U., Pyka, M., Schneider-Hassloff, H., Drexler, R., Krug, A., Sommer, J., Rietschel, M., Witt, S. H., & Kircher, T. (2018). Outgroup emotion processing in the vACC is modulated by childhood trauma and

CACNA1C risk variant. *Social Cognitive and Affective Neuroscience*, *13*(3), 341–348. https://doi.org/10/gcwmnn

Kriegeskorte, N., & Bandettini, P. (2007). Combining the tools: Activation- and information-based fMRI analysis. *NeuroImage*, *38*(4), 666–668. https://doi.org/10.1016/j.neuroimage.2007.06.030

Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(51), 20600–20605. https://doi.org/10.1073/pnas.0705654104

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, *60*(6), 1126–1141. https://doi.org/10/cpdmss

Kubota, J. T., Banaji, M. R., & Phelps, E. A. (2012). The neuroscience of race. *Nature Neuroscience*, *15*(7), 940–948. https://doi.org/10.1038/nn.3136

Lantos, D., Lau, Y. H., Louis, W., & Molenberghs, P. (2020). The neural mechanisms of threat and reconciliation efforts between Muslims and non-Muslims. *Social Neuroscience*, *15*(4), 420–434. https://doi.org/10.1080/17470919.2020.1754287

Lantos, D., & Molenberghs, P. (2021). The neuroscience of intergroup threat and violence. *Neuroscience & Biobehavioral Reviews*, *131*, 77–87. https://doi.org/10.1016/j.neubiorev.2021.09.025

Lau, T., & Cikara, M. (2017). fMRI Repetition Suppression During Generalized Social Categorization. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-04115-8

Levens, S. M., & Phelps, E. A. (2010). Insula and Orbital Frontal Cortex Activity Underlying Emotion Interference Resolution in Working Memory. *Journal of Cognitive Neuroscience*, *22*(12), 2790–2803. https://doi.org/10.1162/jocn.2010.21428

Li, T., Cardenas-Iniguez, C., Correll, J., & Cloutier, J. (2016). The impact of motivation on race-based impression formation. *NeuroImage*, *124*(Pt A), 1–7. https://doi.org/10.1016/j.neuroimage.2015.08.035

Lieberman, M. D., Hariri, A., Jarcho, J. M., Eisenberger, N. I., & Bookheimer, S. Y. (2005). An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. *Nature Neuroscience*, *8*(6), 720–722. https://doi.org/10.1038/nn1465

Lorenzi-Cioldi, F. (1993). They all look alike, but so do we … sometimes: Perceptions of in-group and out-group homogeneity as a function of sex and context. *British Journal of Social Psychology*, *32*(2), 111–124. https://doi.org/10.1111/j.2044-8309.1993.tb00990.x

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Merritt, C. C., MacCormack, J. K., Stein, A. G., Lindquist, K. A., & Muscatell, K. A. (2021). The neural underpinnings of intergroup social cognition: An fMRI meta-analysis. *Social Cognitive and Affective Neuroscience*, *16*(9), 903–914. https://doi.org/10.1093/scan/nsab034

Molenberghs, P., Gapp, J., Wang, B., Louis, W. R., & Decety, J. (2016). Increased Moral Sensitivity for Outgroup Perpetrators Harming Ingroup Members. *Cerebral Cortex (New York, N.Y.: 1991)*, *26*(1), 225–233. https://doi.org/10.1093/cercor/bhu195

Mullen, B., & Hu, L.-T. (1989). Perceptions of Ingroup and Outgroup Variability: A Meta-Analytic Integration. *Basic and Applied Social Psychology*, *10*(3), 233–252. https://doi.org/10.1207/s15324834basp1003_3

Nestor, A., Plaut, D. C., & Behrmann, M. (2011). Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(24), 9998–10003. https://doi.org/10.1073/pnas.1102433108

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLOS Computational Biology*, *10*(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Nissim, N. R., O'Shea, A. M., Bryant, V., Porges, E. C., Cohen, R., & Woods, A. J. (2017). Frontal Structural Neural Correlates of Working Memory Performance in Older Adults. *Frontiers in Aging Neuroscience*, *8*. https://doi.org/10.3389/fnagi.2016.00328

Oosterhof, N., Connolly, A., & Haxby, J. (2016). CoSMoMVPA: Multi-modal multivariate pattern analysis of neuroimaging datain Matlab / GNU Octave. *Frontiers in Neuroinformatics*, *10*. https://doi.org/10.3389/fninf.2016.00027

Ostrom, T. M., & Sedikides, C. (1992). Out-group homogeneity effects in natural and minimal groups. *Psychological Bulletin*, *112*(3), 536–552. https://doi.org/10.1037/0033-2909.112.3.536

Ozker, M., Yu, L., Dugan, P., Doyle, W., Friedman, D., Devinsky, O., & Flinker, A. (n.d.). Speech-induced suppression and vocal feedback sensitivity in human cortex. *eLife*, *13*, RP94198. https://doi.org/10.7554/eLife.94198

Park, B., & Judd, C. M. (1990). Measures and Models of Perceived Group Variability. *Journal of Personality and Social Psychology*, *59*(2), 173–191. https://doi.org/10.1037/0022-3514.59.2.173

Phelps, E. A., Connor, K. J. O., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., & Gore, J. C. (2000). *Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation*. 1–10.

Pickett, C. L., & Brewer, M. B. (2001). Assimilation and Differentiation Needs as Motivational Determinants of Perceived In-group and Out-Group Homogeneity. *Journal of Experimental Social Psychology*, *37*(4), 341–348. https://doi.org/10.1006/jesp.2000.1469

Popal, H., Wang, Y., & Olson, I. R. (2020). A Guide to Representational Similarity Analysis for Social Neuroscience. *Social Cognitive and Affective Neuroscience*, *14*(11), 1243–1253. https://doi.org/10.1093/scan/nsz099

Provins, C., MacNicol, E., Seeley, S. H., Hagmann, P., & Esteban, O. (2023). Quality control in functional MRI studies with MRIQC and fMRIPrep. *Frontiers in Neuroimaging*, *1*. https://doi.org/10.3389/fnimg.2022.1073734

Rai, T. S., Valdesolo, P., & Graham, J. (2017). Dehumanization increases instrumental violence, but not moral violence. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(32), 8511–8516. https://doi.org/10.1073/pnas.1705238114

Riek, B. M., Mania, E. W., & Gaertner, S. L. (2006). Intergroup threat and outgroup attitudes: A meta-analytic review. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *10*(4), 336–353. https://doi.org/10.1207/s15327957pspr1004_4

Ritchie, J. B., Lee Masson, H., Bracci, S., & Op de Beeck, H. P. (2021). The unreliable influence of multivariate noise normalization on the reliability of neural dissimilarity. *NeuroImage*, *245*, 118686. https://doi.org/10.1016/j.neuroimage.2021.118686

Rosch, E., & Lloyd, B. B. (Eds.). (1978). *Cognition and Categorization*. Lawrence Elbaum Associates.

Ruckmann, J., Bodden, M., Jansen, A., Kircher, T., Dodel, R., & Rief, W. (2015). How pain empathy depends on ingroup/outgroup decisions: A functional magnet resonance imaging study. *Psychiatry Research: Neuroimaging*, *234*(1), 57–65. https://doi.org/10.1016/j.pscychresns.2015.08.006

Seber, G. A. F. (1984). *Multivariate Observations* (1st ed.). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470316641

Shen, F., Hu, Y., Fan, M., Wang, H., & Wang, Z. (2018). Racial Bias in Neural Response for Pain Is Modulated by Minimal Group. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00661

Shkurko, A. V. (2013). Is social categorization based on relational ingroup/outgroup opposition? A meta-analysis. *Social Cognitive and Affective Neuroscience*, *8*(8), 870–877. https://doi.org/10.1093/scan/nss085

Smith, E. R., & Zarate, M. A. (1990). Exemplar and prototype use in social categorization. *Social Cognition*, *8*(3), 243–262. https://doi.org/10.1521/soco.1990.8.3.243

Smith, E. R., & Zárate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review*, *99*(1), 3–21. https://doi.org/10.1037/0033-295X.99.1.3

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Stolier, R. M., & Freeman, J. B. (2017). A neural mechanism of social categorization. *Journal of Neuroscience*, *37*(23), 5711–5721. https://doi.org/10.1523/JNEUROSCI.3334-16.2017

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. https://doi.org/10.1002/ejsp.2420010202

Tajfel, H., & Turner, J. (1979). An integrative theory of inter-group conflict. *The Social Psychology of Intergroup Relations*, 56–65.

Tang, H., Ma, G., Zhang, Y., Ye, K., Guo, L., Liu, G., Huang, Q., Wang, Y., Ajilore, O., Leow, A. D., Thompson, P. M., Huang, H., & Zhan, L. (2023). A comprehensive survey of complex brain network representation. *Meta-Radiology*, *1*(3), 100046. https://doi.org/10.1016/j.metrad.2023.100046

Torgerson, W. . S. (1952). Multidimensional Scaling: I. Theory and Method. *Psychometrika*, *17*(4), 401–419. https://doi.org/10.1007/BF02288916

Torgerson, W. . S. (1959). Theory and methods of scaling. New York: John Wiley and Sons, Inc., 1958. Pp. 460. *Behavioral Science*, *4*(3), 245–247. https://doi.org/10.1002/bs.3830040308

Tsantani, M., Kriegeskorte, N., McGettigan, C., & Garrido, L. (2019). Faces and voices in the brain: A modality-general person-identity representation in superior temporal sulcus. *NeuroImage*, *201*, 116004. https://doi.org/10.1016/j.neuroimage.2019.07.017

Tsantani, M., Kriegeskorte, N., Storrs, K., Williams, A. L., McGettigan, C., & Garrido, L. (2021). FFA and OFA Encode Distinct Types of Face Identity Information. *Journal of Neuroscience*, *41*(9), 1952–1969. https://doi.org/10/gnmtf5

Turner, J. C. (2010). *Social categorization and the self-concept: A social cognitive theory of group behavior* (p. 272). Psychology Press.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory* (pp. x, 239). Basil Blackwell.

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2008). The neural substrates of in-group bias: A functional magnetic resonance imaging investigation. *Psychological Science*, *19*(11), 1131–1139. https://doi.org/10.1111/j.1467-9280.2008.02214.x

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2011). Modulation of the fusiform face area following minimal exposure to motivationally relevant faces: Evidence of in-group enhancement (not out-group disregard). *Journal of Cognitive Neuroscience*, *23*(11), 3343–3354. https://doi.org/10.1162/jocn_a_00016

Vanman, E. J. (2016). The role of empathy in intergroup relations. *Current Opinion in Psychology*, *11*, 59–63. https://doi.org/10.1016/j.copsyc.2016.06.007

Verosky, S. C., Todorov, A., & Turk-Browne, N. B. (2013). Representations of individuals in ventral temporal cortex defined by faces and biographies. *Neuropsychologia*, *51*(11), 2100–2108. https://doi.org/10.1016/j.neuropsychologia.2013.07.006

Visser, M., Embleton, K. V., Jefferies, E., Parker, G. J., & Ralph, M. A. L. (2010). The inferior, anterior temporal lobes and semantic memory clarified: Novel evidence from distortion-corrected fMRI. *Neuropsychologia*, *48*(6), 1689–1696. https://doi.org/10.1016/j.neuropsychologia.2010.02.016

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188–200. https://doi.org/10.1016/j.neuroimage.2015.12.012

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Weaverdyck, M. E., Lieberman, M. D., & Parkinson, C. (2020). Tools of the Trade Multivoxel pattern analysis in fMRI: A practical introduction for social and affective neuroscientists. *Social Cognitive and Affective Neuroscience*, *15*(4), 487–509. https://doi.org/10/gg84rn

Weiner, K. S., & Grill-Spector, K. (2010). Sparsely-distributed organization of face and limb activations in human ventral temporal cortex. *NeuroImage*, *52*(4), 1559–1573. https://doi.org/10.1016/j.neuroimage.2010.04.262

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods*, *42*(3), 671–684. https://doi.org/10.3758/BRM.42.3.671

Wilson, J. P., & Hugenberg, K. (2010). When under threat, we all look the same: Distinctiveness threat induces ingroup homogeneity in face memory. *Journal of Experimental Social Psychology*, *46*(6), 1004–1010. https://doi.org/10.1016/j.jesp.2010.07.005

Workman, C. I., Yoder, K. J., & Decety, J. (2020). The Dark Side of Morality—Neural Mechanisms Underpinning Moral Convictions and Support for Violence. *AJOB Neuroscience*, *11*(4), 269–284. https://doi.org/10.1080/21507740.2020.1811798

Yan, Z., Schmidt, S. N. L., Saur, S., Kirsch, P., & Mier, D. (2019). The effect of ethnicity and team membership on face processing: A cultural neuroscience perspective. *Social Cognitive and Affective Neuroscience*. https://doi.org/10.1093/scan/nsz083

Yang, H., Susilo, T., & Duchaine, B. (2016). The Anterior Temporal Face Area Contains Invariant Representations of Face Identity That Can Persist Despite the Loss of Right FFA and OFA. *Cerebral Cortex*, *26*(3), 1096–1107. https://doi.org/10.1093/cercor/bhu289

Yankouskaya, A., Humphreys, G. W., & Rotshtein, P. (2014). The processing of facial identity and expression is interactive, but dependent on task and experience.

Frontiers in Human Neuroscience, *8*, 920. https://doi.org/10.3389/fnhum.2014.00920

Yao, Z., Hu, B., Xie, Y., Moore, P., & Zheng, J. (2015). A review of structural and functional brain networks: Small world and atlas. *Brain Informatics*, *2*(1), 45–52. https://doi.org/10.1007/s40708-015-0009-z

Zaki, J., Schirmer, J., & Mitchell, J. P. (2011). Social Influence Modulates the Neural Computation of Value. *Psychological Science*, *22*(7), 894–900. https://doi.org/10.1177/0956797611411057

10 Statement on own contribution

The entirety of the work presented in this thesis, including planning of the scientific work, data collection, analysis, and interpretation was done completely by me. I am solely responsible for the content of this thesis. I was helped during data collection by research assistants; Aqsa Waris, Dominik Suri, Lennard Schneidewind, and M. Emir Kavukcu. I was also guided during my whole work by the feedback of my first supervisor, PD. Dr. Johannes Schultz. No data from external sources were incorporated, and no experiments, measurements, or materials were generated or processed by third parties.

In preparing this thesis, I used the generative AI tool ChatGPT-4o solely for the purpose of enhancing the readability and fluency of the written text. All content was critically reviewed and edited manually to ensure full alignment with academic standards and originality. I take full responsibility for the final written content of the thesis.

I hereby confirm that my thesis complies with the Statement by the Executive Committee of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) on the Influence of Generative Models of Text and Image Creation on Science and the Humanities and on the DFG’s Funding Activities.

11 Acknowledgments

Big thanks and appreciation to my first supervisor PD. Dr. Johannes Schultz, firstly for giving me the opportunity to study the topic I'm interested in, secondly for being my teacher, tutoring and supporting me through the whole six-year duration of my thesis projects. It has been a pleasure and an honor working with him. Another thank you in order to Prof. Dr. Sebastian Kube for his comments and guidance thought my project. I also wanted to thank my first contributor in this project, Aqsa Waris who helped me kickstart these projects and to my colleagues who helped me acquire the behavioral and fMRI data; Dominik Suri, Lennard Schneidewind, and M. Emir Kavukcu. Also my great appreciation to my ex. and current colleagues: Dr. Daniela Müller, Diana Shih, Dr. Federica Nisini, Dr. Ilinca Serbanescu, Dr. Qëndresa Rramani, and Dr. Xenia Grote who made this journey enjoyable and supported me all along with their advice and care.

This project was fully funded by the Institute of Experimental Epileptoloy and Cognition Research (IEECR, Bonn University Hospitals) directed by prof. Dr. Heinz Beck. Participant recruitment and behavioral pilots were conducted with the help of BonnEconLab, managed by Dr. Holger Gerhardt. Their help has pushed these projects greatly and couldn't have been completed without them. So thank you for your help

12 Appendix

12.1 Self-Assessment Questionnaire (SAQ)

BOGUS PERSONALITY TEST

Your Name:_____

Self-Attributes Questionnaire (SAQ)

This questionnaire has to do with your attitudes about some of your activities and abilities. For the first ten items, you should rate yourself relative to *other college students* your own age by using the following scale:

A	B	C	D	E	F	G	H	I	J
bottom 5%	lower 10%	lower 20%	lower 30%	lower 50%	upper 50%	upper 30%	upper 20%	upper 10%	top 5%

An example of the way the scale works is as follows: if one of the traits that follows were "height", a woman who is just below average in height would choose "E" for this question, whereas a woman who is taller than 80% (but not taller than 90%) of her female classmates would mark "H", indicating that she is in the top 20% on this dimension:

Please rate yourself on the following traits using the scale above.

1. sense of humor _____
2. social skills/social competence _____
3. artistic and/or musical ability _____
4. competency or skill at sports _____
5. physical attractiveness _____
6. leadership ability _____
7. common sense _____
8. emotional stability _____
9. luck _____
10. discipline _____

BOGUS PERSONALITY TEST

Now rate how **certain** you are of your standing on each of the above traits (you may choose any letter).

A	B	C	D	E	F	G	H	I
not at all certain			moderately certain					extremely certain

1. sense of humor _____
2. social skills/social competence _____
3. artistic and/or musical ability _____
4. competency or skill at sports _____
5. physical attractiveness _____
6. leadership ability _____
7. common sense _____
8. emotional stability _____
9. luck _____
10. discipline _____

PLEASE READ

How the scoring works:

All of the ratings are converted into numerical scores and then your ratings on the first part are multiplied by your certainty ratings and then averaged across the ten traits. These scores are then standardized into a single percentile ranking which ranges from 0 to 75. (Standardizing scores allows for easier comparisons across groups of people.)

While you complete the next part of the study, the experimenter will enter the results of this questionnaire into a computer program which will automatically score your questionnaire. As mentioned earlier, the results of this questionnaire will be made available to you and your discussion partner as a means of facilitating the discussion. When you are finished reading, please hand this questionnaire to the experimenter.

12.2 Manipulation Check Questionnaire

Wie hat Sie Ihr SAQ-Ergebnis im Vergleich mit anderen Geisteswissenschaftsstudent*innen fühlen lassen?

sehr anders	etwas anders	leicht anders	neutral	leicht ähnlich	etwas ähnlich	sehr ähnlich

Wie hat Sie Ihr SAQ-Ergebnis über die Gruppe Geisteswissenschaftsstudent*innen fühlen lassen?

Dass Geisteswissenschaftsstudent*innen _____ zu Naturwissenschaftsstudent*innen sind.

sehr anders	etwas anders	leicht anders	neutral	leicht ähnlich	etwas ähnlich	sehr ähnlich

Was war Ihr Ergebnis im SAQ?

Wie haben Sie im Vergleich zum Durchschnitt Ihrer Gruppe Geisteswissenschaftsstudent*innen abgeschnitten?

Ungefähr gleich	Über dem Durchschnit	Unter dem Durchschnitt

Wie war der Durchschnitt von Geisteswissenschaftsstudent*innen verglichen mit dem Durchschnitt von Naturwissenschaftsstudent*innen?

Ungefähr gleich	Sehr unterschiedlich

Was war der Durchschnitt für Geisteswissenschaftsstudent*innen im SAQ?

Was war der Durchschnitt für Naturwissenschaftsstudent*innen im SAQ?

Fanden Sie etwas komisch oder außergewöhnlich an dieser Studie? Wenn ja, was?

Was denken Sie ist das Ziel dieser Studie?