Veit, Susanne; Yemane, Ruta

Working Paper

Judging Without Knowing: How people evaluate others based on phenotype and country of origin – Technical Report

WZB Discussion Paper, No. SP VI 2020-101

Provided in Cooperation with:
WZB Berlin Social Science Center

Suggested Citation: Veit, Susanne; Yemane, Ruta (2020) : Judging Without Knowing: How people evaluate others based on phenotype and country of origin – Technical Report, WZB Discussion Paper, No. SP VI 2020-101, Wissenschaftszentrum Berlin für Sozialforschung (WZB), Berlin

This Version is available at:
https://hdl.handle.net/10419/215833

WZB
Berlin Social Science Center

Susanne Veit
Ruta Yemane

Judging Without Knowing: How people evaluate others based on phenotype and country of origin – Technical Report

Discussion Paper

SP VI 2020–101

April 2020

WZB Berlin Social Science Center

Research Area
Migration and Diversity

Research Unit
Migration, Integration, Transnationalization

WZB Berlin Social Science Center
Reichpietschufer 50
10785 Berlin
Germany
www.wzb.eu

Susanne Veit, Ruta Yemane

Judging Without Knowing: How people evaluate others based on phenotype and country of origin – Technical Report

Discussion Paper SP VI 2020–101

Wissenschaftszentrum Berlin für Sozialforschung (2020)

Affiliation of the authors:

Susanne Veit
WZB Berlin Social Science Center & DeZIM Institut, Berlin, Germany
E.Mail: susanne.veit@wzb.eu / veit@dezim-institut.de

Ruta Yemane
WZB Berlin Social Science Center
E.Mail: ruta.yemane@wzb.eu

Abstract

Judging Without Knowing:
How people evaluate others based on phenotype and country of origin – Technical Report

by Susanne Veit and Ruta Yemane

This report describes the design, data, and main results of an online survey (i.e., the "Judging Without Knowing" survey) that was conducted between October 2017 and June 2018 with more than 2,000 registered members on Clickworker *(*a commercial survey company in Germany). The survey was conducted in order to provide a post-hoc test of the stimulus material (photos) that was used in two correspondence tests on labor market discrimination (i.e., the ADIS and GEMM studies) and to enable further analyses on the role of ethnic stereotypes for ethnic discrimination in hiring. The survey consisted of two parts. The first part of the survey was a post-hoc validation study that aimed at providing an empirical test of the comparability of the photos (phenotype stimuli) from the ADIS and GEMM studies with regard to attractiveness, (ascribed) competence, and sympathy. The second part of the survey studied the stereotypes Germans have about different immigrant groups in Germany. In contrast to previous studies, we asked respondents to rate in how far a range of bipolar adjectives that belong to different stereotype content models (i.e., SCM, 2d-ABC model, and facet model) fit for 38 different ethnic origin groups. In addition, we randomly varied whether respondents had to provide their personal view ("I think …") or their view of the nationally shared stereotype ("Germans think …"). Overall, our findings show that respondents evaluated the photos from the ADIS and GEMM studies differently – but most differences were not substantial. Evaluations differed more strongly between respondents than between photos, and more strongly between photos of males and females and photos series (i.e., original photos and photos that were adjusted with image processing software) than between phenotype groups. The stereotype survey suggests that instruction matters. Respondents rate the different origin groups more positively when asked to express their own opinion than when asked to state the opinion of the Germans. Second, our results raise doubts as for whether Communion is the primary dimension when it comes to stereotypes about immigrant groups in Germany. Ascribed Capacity, Beliefs, and Power seem more important than ascribed Communion. Finally, there seems to be a main divide between the (poor) global south and the (wealthy) global north. Stereotypes about immigrant groups from the global south are generally more negative than stereotypes about immigrants from the global north.

Content

List of Figures

List of Tables

List of Appendix Figures

List of Appendix Tables

Introduction

In the past, extensive research by social psychologists has shown that common beliefs and consensual stereotypes about group specific characteristics do not only affect emotions towards different groups, but also result in discrimination and are (mis)used to legitimize hierarchical intergroup relations and (Agerström & Rooth, 2011; Burgess & Borgida, 1999; Cuddy et al., 2007; Glick & Fiske, 2001; Jost et al., 2005; Jost & Banaji, 1994; Kay & Jost, 2003). They have developed various models that conceptualize stereotypes as the cognitive component of intergroup bias. However, there are differences with respect to the question which stereotype content dimensions are deemed as fundamental (e.g. the *stereotype content model* (SCM, Cuddy et al., 2008; Fiske et al., 2002; Lee & Fiske, 2006), the *facet model* of fundamental content dimensions by Abele et al. (Abele et al., 2016), and the *2d–ABC model* (from here onwards: ABC model; Koch et al., 2016). Yet, the central assumption of all these models is that people do not only perceive and judge others based on their individual and unique combination of traits, characteristics and opinions, but also based on their membership in social groups.

People belong to or are ascribed to many different social groups at the same time (e.g., according to their age, gender, and origin but also according to their professional career or their attractiveness). The present study focuses on the consequences of belonging to a specific *ethnic origin group* (i.e., being an immigrant from or with family roots in different countries of origin) and on the role of *phenotypic appearance* with regard to skin color, hair texture, or facial physiognomy. For this purpose, we draw on a large number of studies on stereotypes about racial, ethnic, and other origin-related minority groups or national groups (Froehlich & Schulte, 2019; Kotzur et al., 2019; Lee & Fiske, 2006; Madon et al., 2001; Phalet & Poppe, 1997). Moreover, there is also empirical evidence suggesting that the way people look plays an important role in how they are perceived and treated by others. Several studies find differences between lighter and darker-skinned minorities with regard to median earnings, net wealth, unemployment, or living in poverty (Castilla, 2008; Painter et al., 2016; Uhlmann et al., 2002).

However, the dynamic behind this finding remains unclear. Does this phenotypic penalty result from the fact that phenotype is a signal of "otherness" and interpreted as a marker of race or ethnic origin – or because evaluations of attractiveness, sympathy, and competence vary systematically between different phenotype categories?

Because modern democracies are characterized by transnational relations and high rates of in- and out-migration, individuals' (ascribed) belonging to national, ethnic, cultural, religious, and racial groups is salient and import. A large and ever-growing number of empirical studies demonstrate that racial, ethnic, and religious minorities and immigrants are treated more negatively than members of dominant societal groups in a wide range of different contexts. Focusing on discrimination based on ethnicity, racial phenotype, and religion, two recent large scale correspondence studies on the German labor market (ADIS: Veit & Yemane, 2018; the German partial study within the GEMM study: Lancee et al., 2019) found evidence for ethnic hierarchies with regard to the likelihood of being invited for a job interview. Correspondence tests are studies in which researchers send out comparable applications from fictitious job candidates to real job openings; these applications vary only the characteristics of interest (e.g. gender, ethnicity) and measure differences in callback rates. Differences in callback rates provide causal evidence of discrimination (for overviews see Gaddis, 2018; Neumark, 2012; Pager, 2007).

The design of the ADIS and GEMM studies is unique as in contrast to the vast majority of previous studies, these two studies allow us to compare employers' responses to applications from second-generation immigrants originating from more than thirty countries, who vary in phenotype and religious affiliation.[1] Both studies find that applicants who themselves or whose parents migrated from poor countries of the global south or from countries with a substantial Muslim population have significantly lower chances of receiving a callback. In addition, the findings point to penalties for phenotypically black and Muslim job applicants, two characteristics that are, however, more likely among the population of the global south than among the population of the global north.

[1] Phenotype and religion varied within the boundaries of plausibility. This means that, for example, applicants of Nigerian origin never applied with a photo showing a person with an Asian phenotype and never signaled a Buddhist affiliation, while applicants with a Chinese background never applied with a photo showing a Black person and never signaled being Muslim.

The *Judging Without Knowing* survey was conducted in order to provide a post-hoc test of the photos that were used in the ADIS and GEMM studies and to enable further analyses on the role of ethnic stereotypes for ethnic discrimination in hiring. Thus, the survey consisted of two parts:

The Photo Survey: The first part of the survey was a post-hoc validation study that aimed at providing a robust and reliable empirical test of the comparability of the photos (phenotype stimuli) from the ADIS and GEMM studies with regard to attractiveness, (ascribed) competence, and sympathy.[2]

The Stereotype Survey: The second part of the survey studied the stereotypes Germans have about different immigrant groups in Germany. In contrast to previous studies on stereotypes in general and German studies on stereotypes in particular, we asked respondents to rate in how far a range of bipolar adjectives that belong to different stereotype content models (SCM: Cuddy et al., 2008; Fiske et al., 2002; facets model: Abele et al., 2016; ABC model: Koch et al., 2016) fit for 38 different ethnic origin groups. In order to add empirical evidence to the discussion of how to best measure stereotype, we decided to randomly vary whether respondents had to provide their personal view ("I think ...") or their view of the nationally shared stereotype ("Germans think ...") (see also Kotzur, Veit, Namyslo, Holthausen, Wagner, & Yemane, 2020).

[2] The photos had been pre-tested prior to the ADIS and GEMM study, but the pre-tests were done with small-n convenience samples.

The "Judging Without Knowing" survey

Research ethics

The research design of the survey was reviewed in advance by the WZB Ethics Committee. Since we asked respondents to evaluate visible minorities on the basis of photos and to judge immigrant groups in a stereotypical manner, the ethics committee demanded that our respondents had the option to refuse answering critical questions, such as the stereotype questions. Thus, we added a "no response" option for virtually all questions.

All survey participants were allowed to leave the study at any time. In addition, we guaranteed their anonymity. Moreover, we informed participants that there were no "correct" or "wrong" answers and that we were aware that it is impossible to evaluate a person based only on a photo, but that we were nonetheless interested in their first impressions, their views, and their thoughts.

The survey was conducted online on a German commercial survey platform. To ensure a sufficiently high share of valid responses and to avoid having respondents "click through" the survey without responding, at least 85% of all questions had to be answered in order to receive the payment code at the very end of the survey. In accordance with the German minimum wage law, participants were paid €2,13 for a survey that took 12 minutes.

Design and implementation

The data collection took place between October 2017 and June 2018. In total, more than 2,000 registered members on *Clickworker* (a commercial survey company in Germany) participated in this study. Quotas were applied to ensure a good distribution across groups, gender and age.

In addition to standard demographic questions, the survey consisted of two parts. In the first part, (I) the photo survey, we asked respondents to evaluate several application photos with respect to "attractiveness," "competence," and "sympathy." In the second part, (II) the stereotype survey, we asked respondents to provide their own stereotypes about several ethnic groups living in Germany by evaluating these groups on semantic differentials with adjective pairs. To explore the role of instruction, we asked half of the sample to state what they believed German stereotypes were about these groups; as it is not clear whether people reproduce the descriptive norms of their society or their own stereotypes (or a mixture of both) when being asked to do indicate what "society thinks" (Brigham, 1972; Stangor & Lange, 1994, Kotzur et al., 2020).

Table 1: Survey overview

Design	Date	I) Photo survey	II) Stereotype survey	Number of participants
0 = Initial survey	October 2017 – March 2018	• 6 photo sets (see Table3) • random sampling within sets, equal assignment probability • n = 6 photos for each participant	• 3 sets of origin groups (see Table 7) • random sampling within sets, equal assignment probability • 3 groups for each participant	n=1,372
Interruption		Mistake in random assignment	Adaptation of design	
1 = Adjusted survey	March 2018 – June 2018	• 7 photo sets • random sampling within sets, different assignment probabilities (dependent on number of observations in initial survey) • n = 7 photos for each participant	• 1 set of origin groups • random sampling, equal assignment probability • 1 group for each participant	n= 969
Total	October 2017 – June 2018			N =2,341

Note, the numbers provided in this table reflect the number of persons who were registered as participants of the survey, but some of them skipped or refused to answer several questions and are therefore omitted from later analyses. For example, 128 persons refused all photo evaluations and 21 persons refused all stereotype evaluations.

Respondent characteristics

In total, 2,341 respondents participated in the survey. Table 2 summarizes their characteristics. The age of participants ranged between 18 and 72, with a mean of 40 years. The gender ratio was balanced, with 50% females and 50% males. On average, every fifth participant was an immigrant or a descendant of an immigrant (18%). Most respondents had either a vocational training certificate (26%) or a diploma or master's degree from university (24%).

Table 2: Sample characteristics

Feature	M (SD) or percent	Min-Max	N
Age	40.31 (10.57)	18-72	2,315
Gender			1,868
female	50%		
male	50%		
other	<1%		
Country of birth			
respondent: Germany (vs. abroad)	92%		2,303
his/her parents: both Germany (vs. one or more abroad)	82%		2,296
Level of education			2,296
general school leaving certificate or lower	11%		
higher entrance qualification	19%		
vocational training (or equivalent)	26%		
Bachelor degree (or equivalent)	14%		
Technician/Master craftsman (or equivalent)	4%		
Master degree(or equivalent)	24%		
PhD or Dr.	2%		

Part 1: Photo Survey

The photo survey aimed at validating the photos that were used in the two field experiments on labor market discrimination (ADIS & GEMM). We tested the photos with respect to perceived *attractiveness*, *sympathy*, and *competence*. The main aim was to provide empirical evidence on the comparability of the photo material in order to gain a better understanding of the role of applicants' phenotypes as a driver of hiring discrimination.

Design & Material

All respondents first read a brief introduction, which informed them that they would see photos that they had to evaluate. They were also informed that there was no "right" or "wrong" answer but that we were interested in their spontaneous opinion and that they could refuse to answer. In the first step, respondents were asked to look at the photos and to answer the following three questions: "How likeable do you find this person on the photo?" (7-point scale, from "not very likeable" to "very likeable"), "How attractive do you find this person?" (7-point scale, from "very unattractive" to "very attractive"), and "How competent does this person appear to you?" (7-point scale, from "very incompetent" to "very competent").

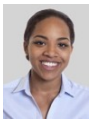
In total, we tested 44 photos (22 photos of males and females, respectively). These photos were used either in the ADIS or the GEMM study (see Table 3). There were three types of photos:

- Adjusted ADIS: First, there were adjusted photos from the ADIS study (in Table 1: sets 1.1 & 1.2). This photo series showed male and female job candidates with red shirts. In order to maximize the comparability between phenotype groups, all eight photos of men and women were based on *one* original photo, respectively, which had been adjusted with image processing software so that it becomes prototypical for one specific phenotype, for example, East Asian, or Southern European White.

- Original ADIS: Second, there were original photos from the ADIS study (sets 2.1 & 2.2). Again, the photo series showed male and female job candidates with red shirts. The photos were only slightly adjusted, so that all males and females had the same upper bod and the same background, and all females had comparable formal hairstyles.

- GEMM: Finally, there were photos from the GEMM study (sets 3.1 & 3.2). This photo series showed male and female job candidates with light blue shirts against a light grey background. All photos were original photos that had been adjusted with an image processing software. Some of the photos were already used in the ADIS study, while others were new. In addition, a new phenotype was added: White 4 (North African).

Table 3: Photos and realized assignments

Pheno-type	Asian 1: East Asian	Asian 2: South-East Asian	Black 1: East African	Black 2: West African	White 1: Central European	White 2: North European	White 3: South European	White 4: North African	
Set 1.1	A1_A_a_f	A2_A_a_f	B1_A_a_f	B2_A_a_f	W1_A_a_f	W2_A_a_f	W3_A_a_f		
ADIS: adjusted photos									
N_{in} N_{ad} N_{total}	0 317 317	0 336 336	0 320 320	0 313 313	1,372 0 1,372	0 338 338	0 314 314		
Set 2.1	A1_A_o_f	A2_A_o_f	B1_A_o_f	B2_A_o_f	W1_A_o_f	W2_A_o_f	W3_A_o_f		FEMALES
ADIS: original photos									
N_{in} N_{ad} N_{total}	307 0 307	186 157 343	365 0 365	331 0 331	13 357 370	18 322 340	152 153 305		
Set 3.1	A1_G_f	A2_G_f	B1_G_f	B2_G_f	W1_G_f	W2_G_f	W2_G_f	W4_G_f	
GEMM									
N_{in} N_{ad} N_{total}	0 330 330	306 0 306	0 337 337	0 304 304	0 308 308	367 0 367	340 0 340	359 0 359	
Set 1.2	A1_A_a_m	A2_A_a_m	B1_A_a_m	B2_A_a_m	W1_A_a_m	W2_A_a_m	W3_A_a_m		
ADIS: adjusted photos									
N_{in} N_{ad} N_{total}	191 112 303	185 115 300	174 120 294	180 99 279	265 16 281	180 100 280	197 84 281		
Set 2.2	A1_A_o_m	A2_A_o_m	B1_A_o_m	B2_A_o_m	W1_A_o_m	W2_A_o_m	W3_A_o_m		MALES
ADIS: original photos									
N_{in} N_{ad} N_{total}	167 153 320	306 0 306	189 94 283	164 174 338	148 160 308	225 51 276	173 154 327		
Set 3.2	A1_G_m	A2_G_m	B1_G_m	B2_G_m	W1_G_m	W2_G_m	W2_G_m	W4_G_m	
GEMM									
N_{in} N_{ad} N_{total}	221 106 327	0 336 336	199 109 308	246 57 303	0 323 323	242 48 290	219 102 321	245 64 309	
ALL N_{total}	1,904	1,927	1,907	1,868	2,962	1,891	1,888	668	

Results

Table 3 provides an overview of the frequency of photo assignments. It differentiates between the total frequency of assignment (N_{total}) and the frequency of assignment in the initial survey (N_{in}) and the adjusted survey (N_{ad}).

Initially, we designed all photos to have the same assignment probability within each set. As the values of N_{in} indicate, however, there was a mistake in the randomization code that led to missing observations (and a strong oversampling of one photo) in most sets. To fill the missing observations, we adapted the survey. Instead of assigning six photos (one out of each series), we sorted all photos into seven groups so that 1) the number of observations for each photo reached about 300 (by distributing the drawing likelihood within each group accordingly) and 2) similar photos were in the same group to avoid repeated exposure (e.g. in Table 3 row 4: W1_A_a_m and row 6: W1_G_m).

All 44 photos were rated on 7-point scales with respect to sympathy, attractiveness, and competence. On average, respondents rated the photos moderately high on sympathy (M=5.30, SD=1.42, see Figure 1), attractiveness (M=4.81, SD=1.47, see Figure 2), and competence (M=5.05, SD=1.31, see Figure 3). The distribution of bars suggests that while all individual photos were positively evaluated (with means larger than 4), adjusted ADIS photos and photos of males generally received slightly more negative evaluations than photos of females and GEMM or original ADIS photos.

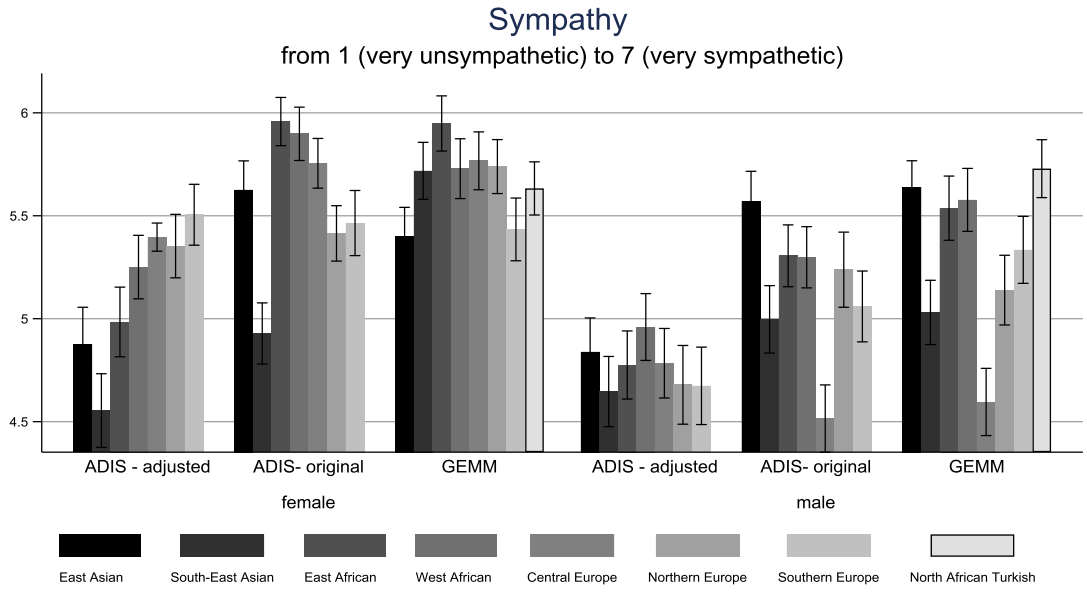


Figure 1: Evaluation of sympathy

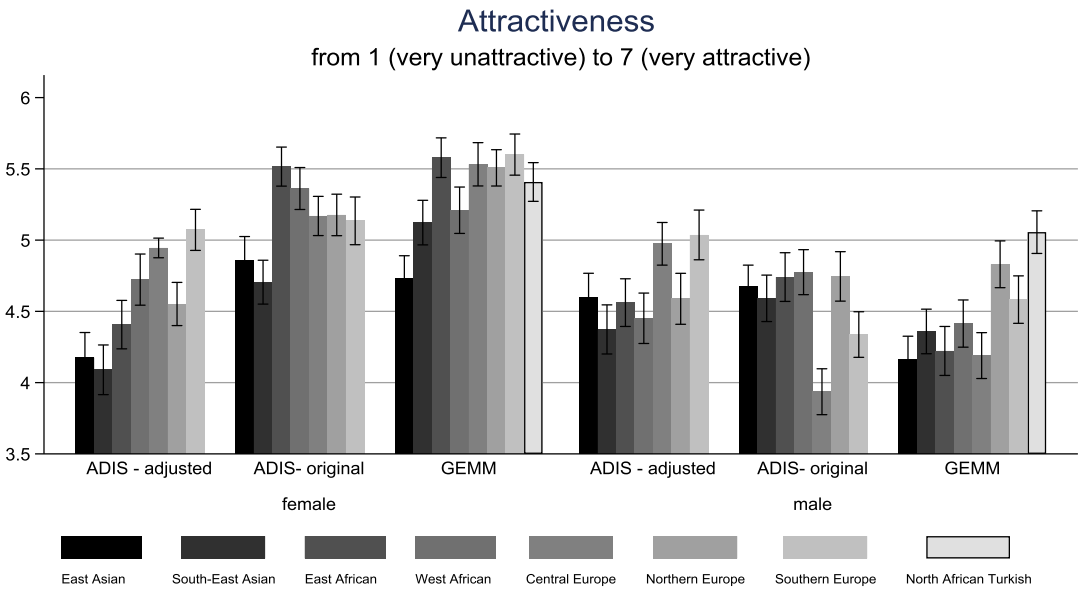


Figure 2: Evaluation of attractiveness

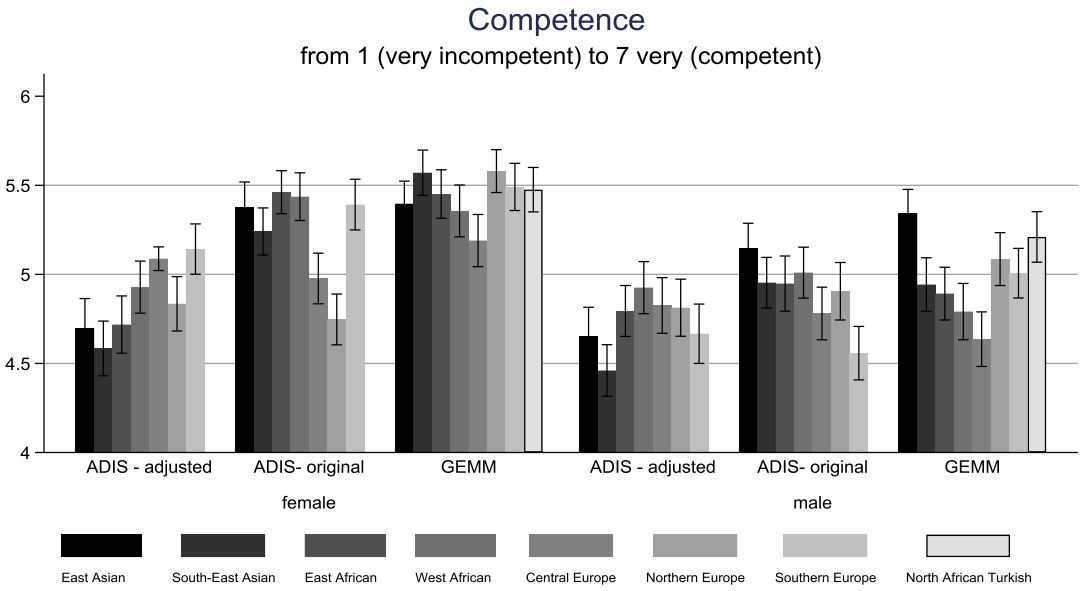


Figure 3: Evaluation of competence

To get a better understanding whether phenotypes matter, we grouped the individual photos to larger phenotype groups (see the photos in Table3: Asian: A1-A2, Black: B1-B2, Northern White: W1-2, and Southern White: W3-W4). In what follows, we show how sympathy, attractiveness, and competence ratings varied between these larger phenotype groups within studies (i.e. ADIS or GEMM) and gender groups (i.e. photos of males or females). Figures 4-6 show bar graphs with confidence intervals for the different phenotype groups. Tables 4-6 provide

the results of linear regression models at the level of observations (m1-m6, respectively) and of linear random slope models with observations nested in individuals for the full sample (m7, respectively). The regression results for single photos (instead of phenotype categories) are provided in the appendix (Tables A1-A3).

Sympathy. Figure 4 illustrates the differences in sympathy ratings by study (i.e., ADIS or GEMM) and gender (i.e., photo of a male or a female person). Respondents rated ADIS photos lower in sympathy than GEMM photos, and males lower than females. As Table 4 shows, some of these differences were statistically significant. Among photos from the ADIS series, sympathy ratings were significantly lower for female Asians compared to the Northern White phenotype, which is the reference category (see Table 4: m1-2). At the same time, sympathy ratings were significantly higher for male Asians with original ADIS photos and GEMM photos compared to the Northern White phenotype (m5-6). Black and Southern White photos were rated significantly more positively than the reference category for females and males with original ADIS photos and males in the GEMM series (m2, m5-6). However, Southern White females in the adjusted ADIS series were rated more negatively than the reference category (m1).

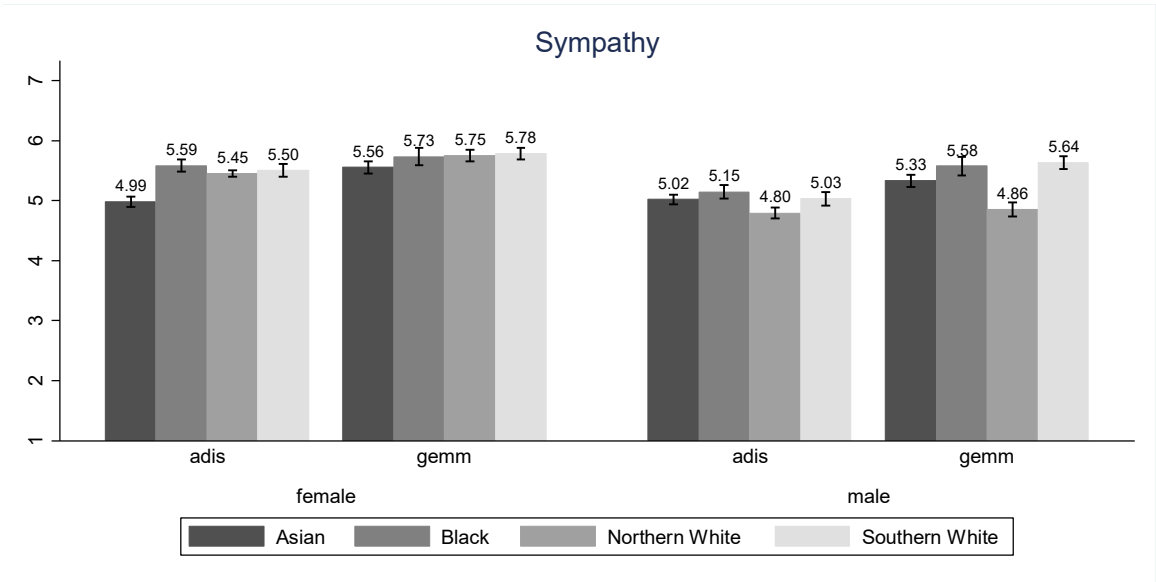


Figure 4: Sympathy evaluation

In the multilevel model for the full sample (m7), all differences were statistically significant, with lower sympathy ratings for Asian photos and significantly higher rating for Black and Southern White photos. Moreover, original photos from the ADIS series and GEMM photos were rated more positively than adjusted ADIS photos. In addition, females were rated more positively than males.

Table 4: Linear regression of sympathy ratings

DV: Sympathy	(1) ADIS adjusted female	(2) ADIS original Female	(3) GEMM female	(4) ADIS adjusted male	(5) ADIS original Male	(6) GEMM male	(7) All
Asian (vs. Northern White)	-.179***	-.126***	-.043	-.001	.121***	.153***	-.031***
	(.0814)	(.0827)	(.0788)	(.0927)	(.0890)	(.0924)	(.0288)
Black (vs. Northern White)	-.026	.061*	.000	.051	.095**	.176***	.064***
	(.117)	(.0938)	(.113)	(.116)	(.112)	(.105)	(.0365)
Southern White (vs. North. White)	-.089***	.098**	-.013	.018	.100***	.225***	.044***
	(.115)	(.0920)	(.0764)	(.116)	(.112)	(.0890)	(.0323)
ADIS original (vs. ADIS adjusted)							.113***
							(.0274)
GEMM (vs. ADIS adjusted)							.168***
							(.0272)
Male (vs. female)							-.157***
							(.0221)
N_{obs}	2143	1588	1640	1399	1476	1652	9898
N_{ind}							1833
R^2	.05	.07	.03	.03	.04	.05	.06

Standardized beta coefficients; Standard errors in parentheses.
Controlled for respondents' age, gender, parents' place of birth, and education (not shown).
Results of linear models (1-6) and linear random intercept models with observations nested in individuals (7).
* $p < .05$, ** $p < .01$, *** $p < .001$

Attractiveness. As Figure 5 illustrates, we also found significant differences between phenotype categories regarding ascribed attractiveness. Attractiveness ratings were much higher and varied much more within female photos than within male photos, with female Asians receiving particularly low ratings. Table 5 confirms that attractiveness ratings were significantly more negative for all female Asians (m1-3) and for male Asians from the adjusted ADIS and the GEMM series (m4, m6) compared to the Northern White reference category. However, male Asians from the original ADIS series (m5) were rated significantly more positively than Northern White males. In addition, Black females from the GEMM study and Black males from the adjusted ADIS series were rated as less attractive than the reference category (m3-4), while Black males from the GEMM study were rated as more attractive (m5). Finally, Southern White females from the adjusted ADIS series were perceived as less attractive than the reference category, while female and male Southern Whites from the original ADIS series received more positive ratings (m2, m5). In the multilevel model for the full

sample (m7), Asians were rated more negatively than the reference category, while Blacks received more positive ratings. In addition, the analysis showed that photos from the GEMM and the original ADIS series were rated more positively than photos from the adjusted ADIS series. Finally, females were considered more attractive than males.

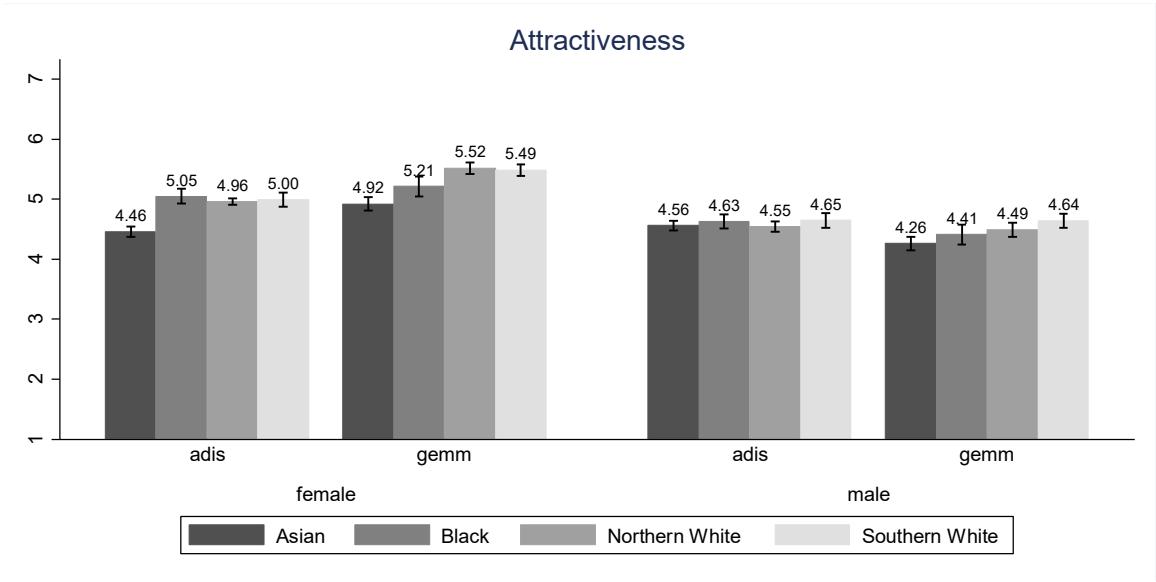


Figure 5: Attractiveness evaluation

Table 5: Linear regression of attractiveness ratings

DV: Attractiveness	(1) ADIS adjusted female	(2) ADIS original Female	(3) GEMM female	(4) ADIS adjusted male	(5) ADIS original male	(6) GEMM male	(7) All
Asian (vs. Northern White)	-.199***	-.140***	-.157***	-.106***	.089**	-.090**	-.092***
	(.0820)	(.0897)	(.0841)	(.0912)	(.0883)	(.0976)	(.0300)
Black (vs. Northern White)	-.035	.020	-.063*	-.097**	.099***	-.029	.003
	(.117)	(.102)	(.121)	(.115)	(.111)	(.112)	(.0381)
Southern White (vs. Northern White	-.098***	.092**	-.018	-.043	.108***	.028	.021*
	(.116)	(.0998)	(.0817)	(.114)	(.112)	(.0939)	(.0337)
ADIS original (vs. ADIS adjusted)							.090***
							(.0286)
GEMM (vs. ADIS adjusted)							.096***
							(.0284)
Male (vs. female)							-.172***
							(.0230)
N_{obs}	2136	1577	1639	1388	1461	1634	9835
N_{ind}							1831
R^2	.06	.11	.05	.04	.06	.03	.07

Standardized beta coefficients; Standard errors in parentheses.
Controlled for respondents' age, gender, parents' place of birth, and education (not shown).
Results of linear models (1-6) and linear random intercept models with observations nested in individuals (7).
* p < 0.05, ** p < 0.01, *** p < 0.001

Competence. In a last step, we analyzed the competence ratings. Figure 6 suggests that there were only small differences between groups. However, Table 6 points to some significant differences between subgroups. Competence ratings were significantly lower for female and male Asians from the adjusted ADIS series (Table 6: m1, m4), while they were significantly higher for female and male Asian from the original ADIS series (m2, m5) and for male Asians from the GEMM series (m6). Black females were generally rated as being more competent than Northern Whites, the reference category (m1-m3), while the ratings for Black males did not differ from the ratings for Northern White males (m4-m6). Likewise, Southern White females from the adjusted ADIS series were rated more negatively (m1) than the reference category, while for Southern White males, we found no difference. The overall pattern differs somewhat from the pattern that we observed for the sympathy and attractiveness ratings (m7). With regard to competence, none of the differences between phenotype groups was statistically significant. Yet, the original ADIS and GEMM photos were again rated more positively than the adjusted ADIS photos. In addition, males were rated significantly more negative than females.

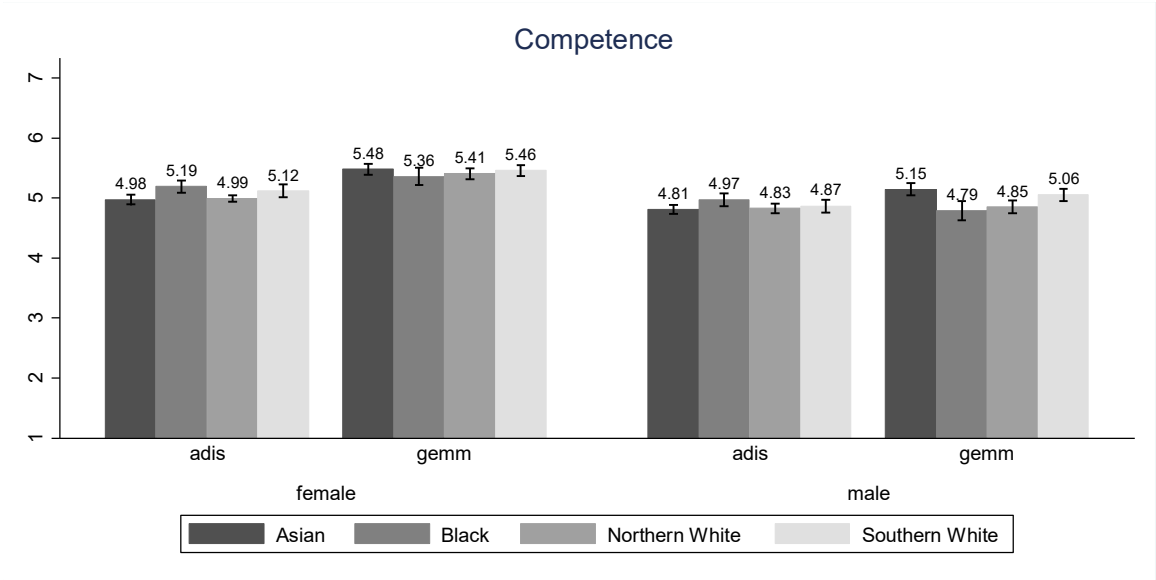


Figure 6: Competence evaluation

Table 6: Linear regression of competence ratings

DV: Competence	(1) ADIS adjusted female	(2) ADIS original Female	(3) GEMM Female	(4) ADIS adjusted male	(5) ADIS original male	(6) GEMM male	(7) All
Asian (vs. Northern White)	-.123***	.166***	.020	-.108***	.088**	.094**	.007
	(.0789)	(.0845)	(.0767)	(.0842)	(.0823)	(.0904)	(.0265)
Black (vs. Northern White)	-.039	.170***	-.065*	.005	.038	-.025	.017
	(.113)	(.0952)	(.109)	(.105)	(.104)	(.103)	(.0335)
Southern White (vs. Northern White	-.104***	.194***	-.003	-.011	.035	.039	.014
	(.112)	(.0936)	(.0746)	(.104)	(.104)	(.0870)	(.0297)
ADIS original (vs. ADIS adjusted)							.089***
							(.0251)
GEMM (vs. ADIS adjusted)							.147***
							(.0249)
Male (vs. female)							-.128***
							(.0202)
N_{obs}	2083	1549	1607	1370	1445	1598	9652
N_{ind}							1802
R^2	.042	.060	.042	.023	.016	.018	.04

Standardized beta coefficients; Standard errors in parentheses.
Controlled for respondents' age, gender, parents' place of birth, and education (not shown).
Results of linear models (1-6) and linear random intercept models with observations nested in individuals (7).
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Covariates and their interaction with photo characteristics. Respondents' age, gender, and origin significantly correlated with the photo evaluations (Appendix table A4: m1-3). Older respondents evaluated photos more positively with regard to sympathy and attractiveness than younger ones. Respondents with foreign roots evaluated the photos significantly more negatively with regard to the competence dimension. Males evaluated the photos generally more negatively on all three dimensions. Respondents' level of education had no effect.

In a next step, we run cross-level interaction models and added interaction terms between respondents' gender and, first, the phenotype on the photo and, second, the gender of the person on the photo to the models with covariates. For the gender-by-phenotype analyses (see Table A5), we found a negative main effects of respondents' gender. In addition, we found negative interaction effects: male respondents judged Asians, Blacks, and Southern Whites in comparison to Northern Whites more negatively than female respondents did on virtually all dimensions. Albeit these interactions were significant in statistical terms, they were very weak in terms of effect size (see Figure 7 for attractiveness).

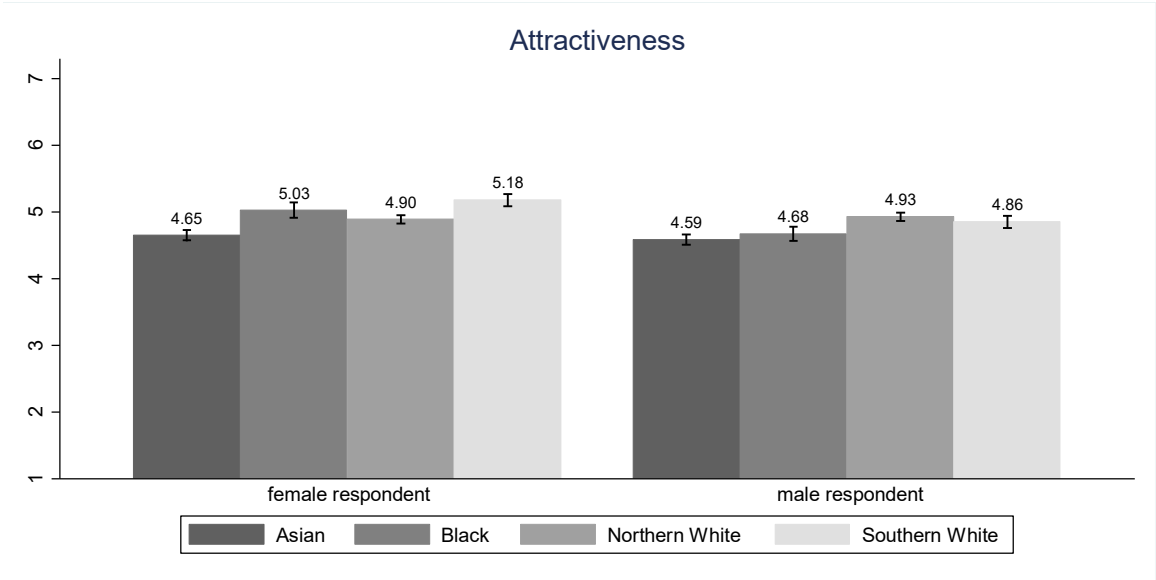


Figure 7: Gender-by-phenotype interaction

The gender-by-gender interaction analyses (see Table A6) revealed that the penalty for male photos in attractiveness and competence evaluations was significantly *less* pronounced among male respondents, even though male respondents tended to give more negative evaluations and male targets tended to receive more negative evaluations. Again, these interaction effects were significant but weak in terms of effect size (for illustration see Figure 8).

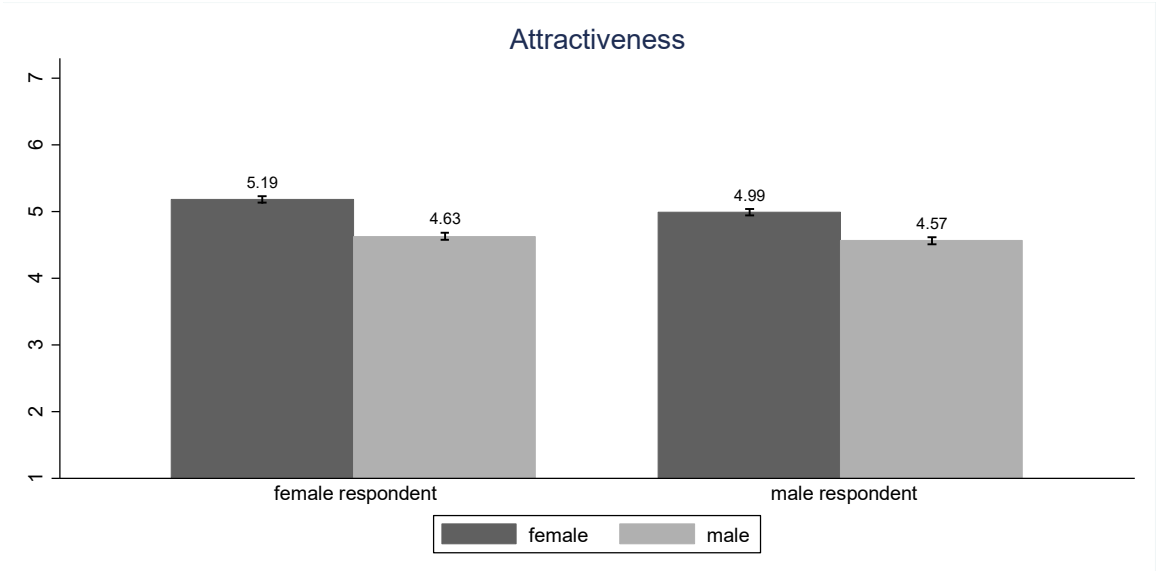


Figure 8: Gender-by-gender interaction

Discussion

In sum, the different photos that we used in the ADIS and GEMM studies were evaluated differently – but most differences were not substantial. Overall, evaluations differed more strongly between photos series (original ADIS, adjusted ADIS, and GEMM) and gender than between phenotype groups (see Tables 4-6: m7, respectively). The only exception are the significantly more negative attractiveness ratings for Asian photos (Table 5, m7). In line with this observation, the comparison between empty regression models with observations nested in photos (N=44) and models with observations nested in individuals (N≈2,300) suggested that ratings vary more strongly between respondents (ICC_{sym}=.36, ICC_{attr}=.36, ICC_{comp}=.42) than between photos (ICC_{sym}=.08, ICC_{attr}=.09, ICC_{comp}=.05).

Most importantly, phenotypes that are typically associated with low status, disadvantages, and discrimination, i.e. Black phenotypes, were not rated more negatively. Black photos were rated just as positively as Northern Whites in terms of attractiveness and competence. They were also rated as more likeable than Northern White phenotypes. Asian photos, by contrast, were rated as less likeable and less attractive than Northern Whites, while they were perceived similar to Northern Whites with regard to competence. This observation is in line with the stereotype of Asians, who are often portrayed as being cold but competent (Cuddy et al., 2008; Lee & Fiske, 2006). For Blacks, by contrast, US studies suggest that they are often perceived as being low in warmth and competence (Devine & Elliot, 1995; Dovidio et al., 1986; Fiske, 2018). Given the negative stereotype about Blacks the photos of Black people were evaluated more positively than expected. One possible explanation for this result is that the data was collected in Germany, where stereotypes about Black are probably weaker than in the U.S. (but see Kotzur et al., 2019, Samples, 2019).

Part 2: Stereotype survey

The second part of the survey measured the stereotypical views Germans have about two groups; namely, either about other Germans or about various immigrant groups in Germany. More specifically, we tested how respondents evaluate different immigrant groups in Germany with regard to a range of various descriptive adjectives.

Design and Material

Respondents first read an introduction (see Figure A2), which informed them that they will be asked to evaluate three (and later one) randomly assigned groups of people living in Germany on a list of 15 adjectives (for the instructions in German, see Appendix Figure A3). They were then asked to evaluate the German language skills of different ethnic groups and the extent to which different ethnic groups are similar to Germans. They were also asked how certain they felt about their evaluation (i.e. stereotype strength). With the first question, we introduced the perspective of evaluation by either asking respondents what *they personally think* or what *Germans think* about various social groups in Germany. We varied the perspective between the respondents but kept it constant for individual respondents.

After this, the evaluation started. Before we encountered the aforementioned randomization problem, we asked each respondent to evaluate three out of 38 origin groups (see Table 7 below). The specific target groups were chosen because they were used in either the ADIS study or the GEMM study. The social groups were randomly assigned out of three blocks (see the first column in Table 7). After we encountered the randomization problem (see chapter I), we changed the design so that only one ethnic origin group was assigned out of the total pool with 38 groups. The assigned social group was named at the top of each page (e.g.: "Romanian immigrants living in Germany"). In addition, a map of the world appeared at the top of the screen. On this map, the respective country of origin was highlighted.

Below the map, each respondent saw 15 pairs of descriptive adjectives (e.g. "dominated" and "dominating") which they had to rate on a 7-point scales (see Table 8 and Figure A6 for all item pairs in German). We asked respondents to evaluate the assigned social group on these semantic differentials – either by providing their own opinion or by indicting what Germans think about this

group. The 15 adjective pairs were presented in random order, and they were followed by three additional questions concerning groups' similarity with Germans, their German language skills, and respondents' certainty of evaluation, as an indicator of stereotype strength (again, see Table 8).

Table 7: Origin groups

Block	Country of origin	Perspective "self"	Perspective "Germans"	Total Freq.
1	Germany	66	65	131
	Turkey	61	77	138
	Bulgaria	73	59	132
	France	63	59	122
	Greece	62	52	114
	Italy	60	59	119
	Netherlands	63	74	137
	Norway	58	63	121
	Poland	57	66	123
	Romania	67	68	135
	Spain	65	66	131
	Switzerland	77	76	153
	United Kingdom	62	77	139
2	Albania	64	74	138
	Bosnia and Herzegovina	62	66	128
	Macedonia	57	50	107
	Russia	60	83	143
	Egypt	73	66	139
	Iraq	61	68	129
	Iran	69	71	140
	Lebanon	61	67	128
	Morocco	50	61	111
	Ethiopia	62	65	127
	Nigeria	70	71	141
	Uganda	64	68	132
	South Africa	58	56	114
3	China	69	86	155
	Dominican Republic	75	71	146
	Indonesia	73	67	140
	India	54	77	131
	Japan	61	70	131
	Malaysia	74	74	148
	Mexico	74	69	143
	Pakistan	59	78	137
	South Korea	81	86	167
	Trinidad and Tobago	85	56	141
	USA	54	63	117
	Vietnam	58	75	133
N_{obs}		2,462	2,599	5,061

We based the selection of descriptive adjectives on three sources: First, the stereotype content model (SCM: Cuddy et al., 2008; Fiske, 2018; Lee & Fiske, 2006), second, the facet model of fundamental content dimensions by Abele and colleagues (2016), and third, the ABC model (Koch et al., 2016).

According to the stereotype content model (SCM), WARMTH and COMPETENCE are the two fundamental stereotype content dimensions. SCM studies often use one-dimensional scales (Fiske et al., 2002: „As viewed by society, how competent are members of this group?“) to measure stereotype content. In some studies, only *one* item per dimension was presented (e.g. “warm” and “competent” in Lee & Fiske, 2006), but in most studies several items were used. Typical items or descriptions used in SCM studies are ‘*warm*’, ‘*benevolent*’, ‘*likeable*’, ‘*trustworthy*’, ‘nice’, ‘friendly’, and ‘sincere’ for WARMTH and ‘*competent*’, ‘*laborious*’ ‘*reliable*’, ‘*highly educated*’, ‘skillful’, and ‘able’ for COMPETENCE (see e.g. Cuddy et al., 2008). The items that we used in our own study are highlighted in italics.

The ABC model differentiates between AGENCY, progressive BELIEFS, and COMMUNION. In a study with a German sample, Koch et al. (2016) presented their items on semantic differentials. However, they did not present the items separately (i.e., one after another) but in three blocks (one for each dimension), and they asked respondents to judge several social groups on each dimension. They used the following item blocks to measure their three stereotype content dimensions (here we only mention one pole of the semantic differential): A) AGENCY: ‘*high in status, dominant*, *confident*, rich, powerful, competitive’; B) BELIEFS: ‘*traditional, religious, conservative*, conventional’; and C) COMMUNION: ‘*trustworthy, likable, benevolent, warm*, sincere, altruistic’. From each item block, we included three to four adjectives in our analyses. The items are again highlighted in italics.

Finally, Abele and colleagues (2016) proposed a facet model of stereotype content that differentiates between ASSERTIVENESS (AA) and COMPETENCE (AC) as facets of agency and between WARMTH (CW) and MORALITY (CM) as facets of communion. To measure these four facets, Abele and colleagues (2016) presented several adjectives on five-point scales, some of them being more similar to one-dimensional scales (e.g. from “not capable” to “very capable”) and others being more similar to semantic differentials with bipolar adjectives (e.g. from “very cold in relations with others” to “very warm in relations with others”). In total, they used twenty adjective pairs (again, only one pole is mentioned here): CW – “very caring”, “very *warm* in relations with others”, “very empathetic”, “very

affectionate" and "very friendly"; CM – "just", "very fair", "very considerate", "very *trustworthy*", and "very *reliable*"; AA – "very *self-confident*", "stands up well under pressure", "never gives up easily", "has leadership qualities" and "feel very superior"; and AC – "very efficient", "very capable", "very *competent*", "very intelligent" and "very clever". Again, the items that we used in the present study are highlighted in italics.

For our own study, we decided to combine all three strategies. We used semantic differentials with 15 pairs of polar adjectives at the opposite ends of 7-point scales (see Table 8 below). The 15 adjective pairs were presented in random order, and they were followed by three additional questions concerning the groups' similarity with Germans, their German language skills, and respondents' certainty of evaluation, as an indicator of stereotype strength. Moreover, respondents were asked to indicate either their own or Germans' stereotypes about the respective group.

Table 8 in the Results section lists the positive value of all 15 adjective pairs, sorted by the three major content dimensions that emerge from SCM, the facet model, and the ABC-model. The enclosed superscripts next to the adjectives indicate whether and from which stereotype content model each adjective was taken or whether it was self-generated by the authors.

Results

Table 8 below provides the summary statistics for all 15 descriptive adjective pairs and the three additional items measuring similarity, language skills, and stereotype strength averaged across all origin groups. In Table 8 we separated the ratings by the two perspectives "self" or "Germans". Overall, evaluations were moderately positive: most evaluations were on average close to the theoretical midpoint of the scale.

However, evaluations varied considerably between respondents who had been asked to provide their own stereotypes and respondents who had been asked to indicate what Germans think about different immigrant groups in Germany (see the last columns in Table 8). Respondents' own opinion was more positive for all adjective pairs. A MANOVA confirmed the statistical significance of the differences between "perspective" groups: Roy's largest root=.0306 and Wilks' lambda=.0970 $F(18,4147)=7.06$, $p<.001$. There were also significant differences in the evaluation of similarity, with higher similarity ratings when providing one's

own option than when providing Germans' views of the different origin groups: t(4857)=6.1003, p<.001 (see Figure 9). With respect to stereotype strength, however, the opposite pattern emerged (see Figure 10). Respondents were on average quite confident about their evaluations, and this confidence was even higher among participants who responded on behalf of Germans: t(5044)=-2.4036, p<.01.

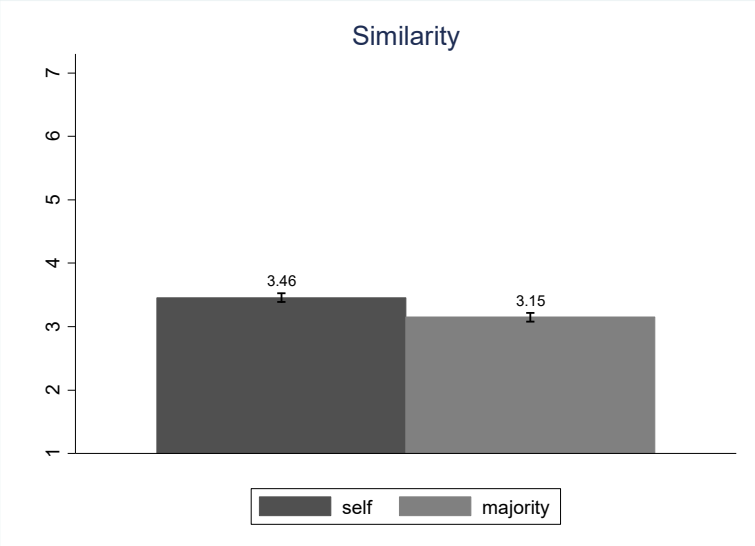


Figure 9: Similarity by perspective

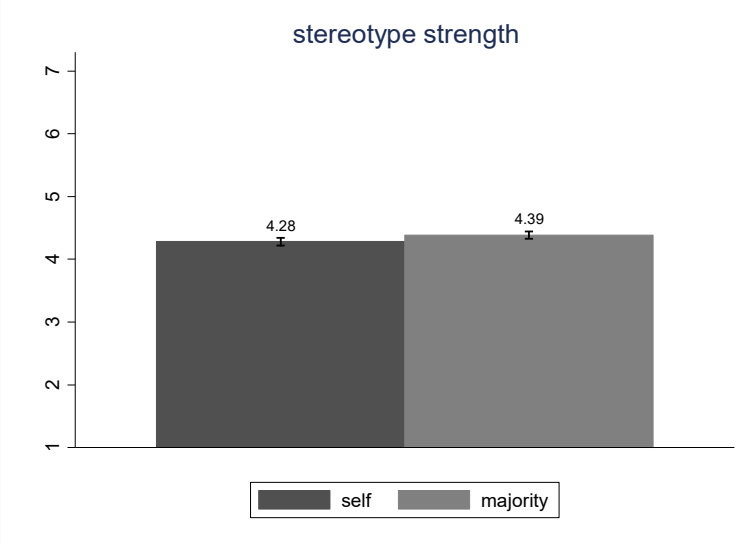


Figure 10: Stereotype strength by perspective

Table 8: Summary statistics of adjective

Dimension	Facet		Total			Perspective "self"			Perspective "Germans"		
		Items	N	mean	sd	n	mean	sd	n	mean	sd
COMMUNION	MORALITY WARMTH	Trustworthy [1,2,3]	4882	4.215895	1.53898	2358	4.407125	1.443779	2524	4.037242	1.602676
		Benevolent [12]	4892	4.351186	1.567317	2380	4.528571	1.483984	2512	4.183121	1.624896
		Reliable [3]	4851	4.294991	1.543251	2337	4.438169	1.453226	2514	4.161893	1.611379
		Likeable [12]	4921	4.504369	1.432361	2381	4.635447	1.351558	2540	4.381496	1.494025
		Warm [1,2,3]	4875	4.718359	1.404628	2370	4.810127	1.361373	2505	4.631537	1.439253
AGENCY	COMPETENCE ASSERTIVENESS	Laborious [2]	4867	4.590507	1.524892	2349	4.737335	1.408924	2518	4.453535	1.613912
		Highly educated [4]	4911	4.158623	1.535157	2373	4.304678	1.441191	2538	4.022065	1.606398
		Competent [2,3]	4839	4.393056	1.443035	2341	4.516873	1.349673	2498	4.277022	1.51647
		Successful [3]	4882	4.284924	1.454118	2359	4.419245	1.356411	2523	4.159334	1.529507
		High status [1]	4882	3.897173	1.53842	2359	4.031793	1.457136	2523	3.771304	1.601576
		Dominating [1]	4796	4.167223	1.411844	2325	4.221935	1.353182	2471	4.115743	1.463302
		Self-confident [1,2,3]	4891	4.607851	1.46896	2369	4.655129	1.402405	2522	4.563442	1.527787
BELIEFS		Traditional [1]	4953	3.364224	1.738971	2392	3.454431	1.681826	2561	3.279969	1.786917
		Religious [1]	4910	3.488187	1.774334	2384	3.556208	1.73029	2526	3.42399	1.812925
		Conservative [1]	4865	3.486125	1.577058	2349	3.490847	1.526463	2516	3.481717	1.623162
Similarity		very similar[ab]	4859	3.299239	1.770074	2360	3.458051	1.747989	2499	3.14926	1.778049
German Language Skills		very good[ab]	4821	3.576644	1.723461	2336	3.747003	1.697647	2485	3.416499	1.732507
Stereotype Strength		very certain[b]	5046	4.336901	1.540807	2456	4.283388	1.563426	2590	4.387645	1.517607

[a] These items were not shown on screens where respondents were asked to evaluate "Germans".
[b] These items were always presented last (fixed order).
[1] Items belong to ABC-model; [2] Items belong to SCM.; [3] Items belong to facet model.; [4] Item is self-generated.

Factor structure: Stereotype content dimensions. The intraclass correlation coefficients (ICC) resulting from empty regression models with evaluations of the fit of the descriptive adjectives as dependent variable (measured at the level of observations) and origin groups as units at the second level were moderate to high (.08 < ICC < .37), which suggests that the ethnic target group matters.[3] To explore the dimensional structure of the data, we therefore used a two-level explorative factor analyses (MEFA) in Mplus (with oblique rotation and ratings nested in origin groups).

At both levels (i.e., the level of observations and the level of origin groups), three factors with eigenvalues greater one emerged (see Table 9). At the within level, a fourth factor with an eigenvalue of .93 was confirmed. At the between level, the eigenvalue of the fourth factor equaled .52. As appendix Table A7 shows, however, none of the models with two, three, or five factors at the between level converged. In addition, the fit of four-factor-models at both levels was good.[4] It was even better than any other solution with one to five factors at the within and/or between level for which fit indices could be calculated (except for the model with five within and four between factors, see Tables A7-A8). In the four factor solution, most items loaded clearly on one factor, except for "trustworthy" (within and between level), "likable" (between level) and "benevolent" (between level), which had substantial cross-loadings (see Table 9)[5]. Overall, the pattern of loadings only partly met the propositions of the SCM, the facet model, and the ABC model, respectively.

At the *within level*, the first factor combined items measuring competencies in SCM and the facet model with the status-item from the agency dimension of the ABC model, and trustworthiness and reliability, two items that are considered to measure morality in the facet model. However, it is easy to think of trustworthiness and reliability as important qualities in the work context, which implies a conceptual closeness to the competence dimension. Likewise, the close link between competence and status ties in with SCM's proposition that perceived status is an important predictor of and thus highly correlated with

[3] Empty regression models confirmed that there is substantial variation between origin groups in the full sample but also in the two subsamples: total: .08 (warm) <ICC<.37 (status); "self": .07(warm) <ICC<.35(status); "Germans": .10 (warm)<ICC<.40 (status).

[4] The chi-square tests were significant, but chi-square test tend to "reject reasonably specified models as a result of large sample sizes" (Huang & Cornell, 2016, S.7).

[5] Excluding these items from the analysis, however, made other items cross-loading. Moreover, at the between level only one trait (warm) from the fourth factor (*Communion*) would remain.

competence stereotypes. Thus, all the items that loaded on the first factor indicate whether an individual is able to reach his or her goals and his or her quality as a team member (by being reliable and trustworthy). We therefore called this first factor *Capacity*. On the *second factor* loaded all items of ABC model's beliefs dimension. We therefore also named this dimension *Beliefs*. The *third factor* included only "dominant" and "self-confident" – two items that measure agency in the ABC model. Since these two items do not cover the status aspect of agency but only the power aspect, we called this factor *Power*. Finally, the fourth factor covered communion items, two from the warmth ("benevolent" and "trustworthy") and two from the morality facet ("warm" and "likable").

Table 9: MEFA – factor loadings

		Within				**Between**			
		1: Capacity	2: Beliefs	3: Power	4: Commu-nion	2: Capacity	3: Beliefs	1: Power	4: Comm-union
	Eigenvalue	6.85	1.50	1.30	.93[1]	1.68	1.45	11.17	.52[1]
Items	Factor loadings								
1	Competent	.739				.983			
2	Laborious	.654				1.092			
3	Reliable	.715				.935			
4	Educated	.794				.990			
5	High in Status	.688				.772			
6	Successful	.781				.963			
7	Trustworthy	*.504*			*.443*	.721			
8	Modern		.789				.679		
9	Secular		.635				.876		
10	Liberal		.660				.756		
11	Dominant			0.544				1.028	
12	Self-confident			0.730				.802	
13	Warm				.727				1.135
14	Likeable				.658	*.521*			*.651*
15	Benevolent				.544	*.580*			*.500*

Note: Factor loadings smaller than .40 are not shown.
For each item, the highest factor lading is highlighted in bold. Items with substantial cross-loading are highlighted in italics.
[1] The fourth factors at the within- and between-level are included despite their low eigenvalues, because none the models with three factors at the between level converged.

At the between level, a very similar pattern emerged (again, see Table 9). Note, however, while the content of the four factors was very similar, the sorting of factors in terms of eigenvalues was considerably different between levels. At the within level, *Capacity* was the strongest factor, but at the between level *Power* had by far the strongest eigenvalue. Moreover, at the between level *Communion* was more difficult to confirm, because three out of four items had considerable cross-loadings and the factor had an eigenvalue below one.

Since at both levels three factors with eigenvalues larger than one emerged, in a next step we run for each origin group separate factor analyses with maximum three factors to be retained. Table 10 illustrates the emerging factor structure. Items that loaded on the same factor are shown in similar color. Items that loaded on two factors are shown have a split cell with two colors. Negative loadings are indicated by means of a hyphen. Loadings below .40 are identified by the word "none".

With some exceptions, the following pattern emerged: The *Beliefs* dimension was confirmed for a vast majority of origin groups (see the last three columns in Table 10). There was also surprisingly high consensus with regard to the *Power* dimension. For *Capacity* and *Communion* items the pattern of results was somewhat mixed. For a relatively high number of origin groups we found that items from both dimensions loaded on one and the same factor, which suggests that they measure the same latent construct. This observation fits to the cross-loadings of the "communion adjectives" in the between–level results of the MEFA in Table 9.

Table 10: Factor loadings in separate factor analyses by origin groups

Origin group	competent	laborious	reliable	trustworthy	educated	successful	high in status	dominant	self-confident	benevolent	sympathy	warmth	traditional	religious	conservative
Germany															
Turkey							none								
Bulgaria												-			
France															
Greece															
Italy								-							
Netherlands															
Norway															
Poland															
Romania															
Spain															
Switzerland								-							
United Kingdom															
Albania															
Bosnia & Herzegovina															
Macedonia															
Russia							none								
Egypt															
Iraq															
Iran															
Lebanon															
Morocco															
Ethiopia															
Nigeria												-			
Uganda															
South Africa															
China														none	-
Dominican Republic															
Indonesia															
India															
Japan								-							
Malaysia															
Mexico															
Pakistan															
South Korea															
Trinidad and Tobago								--							
USA					none										
Vietnam															

Note: Table 10 shows the factor structure that emerged in principal-component factor analyses with a maximum number of three factors to be retained and oblique rotations (in STATA: promax). Items that loaded on the same factor are shown in the same color. Factor loadings smaller than .40 are identified by the word "none". Hyphens indicate negative factor loadings.

Factor scores as indicators of ethnic stereotype. Based on the factor structure that emerged in the factor analyses, we computed indices for *Capacity*, *Power*, *Beliefs*, and *Communion* by averaging the evaluations across all items that belong to each of the four stereotypes content dimensions (see Table 11).[6] Because of the centrality of *Capacity* and *Communion* in the SCM and the facets model, we distinguished between these two dimensions despite the partly mixed factor analyses results.

The reliability of the resulting scores did not vary with the perspective of rating (i.e., "self" vs. "Germans"). Moreover, the stereotype scores were all around the theoretical midpoint of the scale and they were all positively correlated ($p<.001$, respectively); with particularly strong correlations of *Capacity* with *Beliefs*, on the one hand, and *Communion*, on the other hand (see Table 11). The correlation of *Power* with *Beliefs* and *Communion* was in comparison rather low ($r=.29$ and $r=24$, respectively), but still highly significant.

Table 11: Stereotype content scores

	Total									**"Self"**			**"Germans"**		
Stereotype scores	N items	N obs	M	sd	α	**Correlations** Ca	Po	Be	Co	M	sd	α	M	sd	α
Capacity competent, laborious , reliable, trustworthy, educated, successful, high in status	7	5,00	4.26	1.31	.95	1				4.40	1.20	.94	4.13	1.39	.95
Power dominant, self-confident	2	4,95	4.39	1.25	.64	.38	1			4.44	1.18	.62	4.34	1.31	.65
Beliefs traditional, religious, conservative	3	5,00	3.45	1.47	.82	.60	.29	1		3.51	1.42	.81	3.40	1.51	.82
Communion benevolent, likeable, warm	3	4,99	4.52	1.27	.82	.69	.24	.47	1	4.66	1.20	.82	4.40	1.31	.83

[6] We are aware of the discussion concerning the question whether stereotype content can be measured as simple scale means when investigating stereotypes about different target groups, or whether researchers need to apply a latent variable framework and to establish measurement invariance (e.g., Kotzur et al., 2019: Kotzur et al, 2020). However, since this report only aims at proving information about the data collection and at illustrating potential applications of the survey results, we decided for the most simple and easy-to-understand procedure and computed mean scores.

Since we were primarily interested in the stereotypes people in Germany have regarding different origin groups, in the next step we explored how these stereotype scores differed between origin groups. We first explored the amount of variation between origin groups by means of empty regression models with stereotype content scores as the dependent variable and observations nested in origin groups (see Table A10). The resulting intraclass correlation coefficients suggested that there is more variation in groups' ascribed *Beliefs* and *Capacity* than there is in ascribed *Power* and *Communion*. Figure 11 illustrates how the different origin groups scored on all four dimensions. With respect to *Capacity* and *Communio*n, our results were very similar to the pattern reported by Froehlich and Schulte (2019). Respondents rated Germans very highly on *Capacity, Beliefs*, and *Power* and lower on *Communion.* From the SCM perspective, this makes intuitive sense, since Germans are the in-group and in-groups are usually perceived to be "warm and competent", while Germans as an origin group are often stereotyped in an ambivalent way, e.g. "competent but cold". Moreover, the patterns of results suggest that Germans and immigrant groups from Western democracies were perceived to be high on all four stereotype content dimensions. The stereotype of Eastern Asians was very similar to that of Westerners, but they were described as having less power. Finally, immigrants from the global south were described as being low with regard to all four dimensions (i.e. as rather incapable, traditional, powerless, and low in communion).

Finally, to illustrate the role of ethnic stereotypes, we run regression models. For each origin group we computed average stereotype content values, stereotype strength, and similarity ratings. Based on these average values, we run a simple linear regression of average similarity on average *Capacity, Beliefs*, *Power, Communion* and stereotype strength (N=37)[7]. This very straightforward approach explained 92% of the variance in origin groups' similarity to Germans, with significant positive regression coefficients for average *Power* (b=.55, se=.17, p<.01), *Beliefs* (b=.67, se=.13, p<.001) and stereotype strength (b=.77, se=.27, p<.01). The regression coefficients of average *Capacity* and *Communion* were not significant.

[7] Here we considered only immigrant groups (N=37), because for Germans as target group we did not ask respondents to evaluate their similarity to Germans.

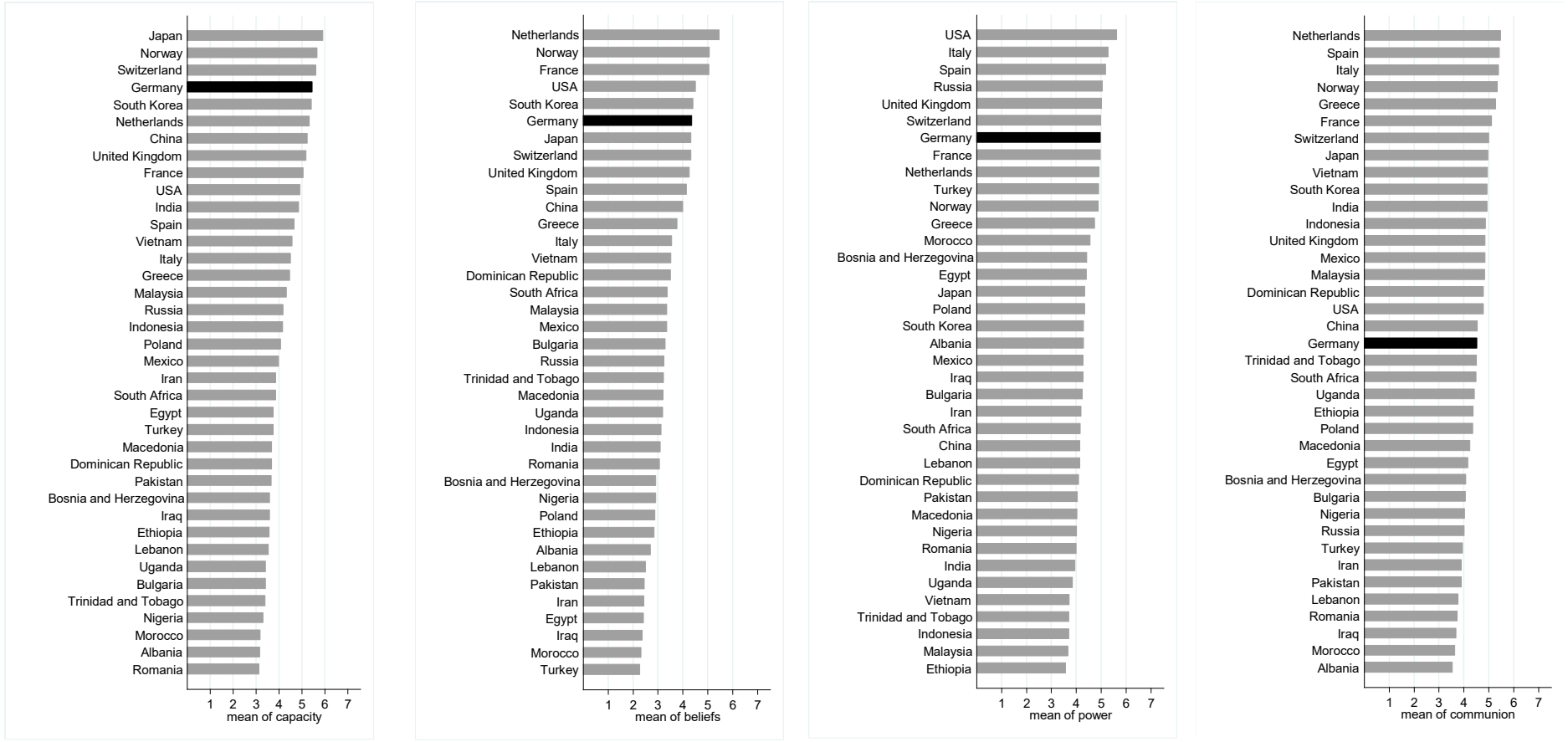


Figure 11: Capacity, beliefs, power, and communion scores of origin groups

When instead specifying a much more complex cross-classified multilevel regression model with controls and many more observations (see Table 12), we observed the same trend: stereotypes about immigrant groups' power and beliefs were (among the four stereotype content dimensions) the most powerful predictors of perceived similarity. In addition, stereotype strength mattered considerably: the more confident respondents felt about their stereotypes regarding a certain immigrant group, the higher the likelihood that this groups was perceived to be rather similar to Germans. Finally, in this analysis we again confirmed the impact of the perspective of evaluation. Similarity ratings were generally higher among respondents who responded on behalf of Germans in general.

Table 12: Cross-level regression of similarity

DV: similarity	b	se	p>|z|
ORIGIN GROUPS			
Stereotype scores:			
Capacity	-.159	(-.98)	.326
Power	.557***	(3.37)	.001
Beliefs	.663***	(4.97)	.000
Communion	.372*	(2.18)	.029
Stereotype Strength	.852**	(3.20)	.001
RESPONDENTS			
Age	-.004	(-1.73)	.083
Gender (ref.: female)			
Male	.035	(.66)	.508
other	-.317	(-.31)	.758
Migration background (ref.: no)			
Yes	-.180*	(-2.56)	.010
Level of education (ref.: low)			
Medium	.110	(1.30)	.193
High	.133	(1.49)	.136
Perspective of evaluation (ref.: "Self")			
Germans	-.330***	(-6.31)	.000
Constant	-5.839***	(-7.35)	.000
RANDOM EFFECT PARAMETER	Estimate	Std. Err.	95% Conf. Interval
_all: var (R.origin group)	.084	(.023)	.049-.143
Respondents: var(constant))	.594	(.043)	.516- .685
Var(Residuals)	1.365	(.038)	1.292- 1.443
$N_{observation}$	4,341		
$N_{respondents}$	1,806		
$N_{origin\ groups}$	37		

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Model specification in STATA: mixed y x_1-x_n || _all: R.origin-group || respondents:
Note: No ratings of Germans as target group are included, because for this target group we have no observations on the dependent variable.

Discussion

When judging others, people often draw on stereotypes. Previous research has developed different models of stereotype content dimensions. Yet, none of these models has specifically focused on different origin groups. This study explored how immigrants groups in Germany are typically evaluated with respect to several descriptive adjective that reflect progressive beliefs, communion, and agency (or facets of the latter).

There are three main findings:

1) Instructions matter: When being asked to provide their own opinion, respondents rated the different origin groups more positively than when being asked to indicate the view of Germans.

2) Communion is not the primary dimensions when it comes to stereotypes about immigrant groups in Germany: In this study, we took into account a set of bipolar adjectives that reflect the different stereotype content dimension proposed by SCM, the ABC-model and the facet model. The results of a multilevel explanatory factor analysis were partly compatible with all three models. Four factors, and thus four stereotypes content dimensions, emerged*: Capacity, Beliefs, Power* and *Communion.* However, *Communion* – which is the primary dimension in SCM – is the dimension that received the least empirical support. When predicting origin groups' similarity to Germans in regression models by group stereotypes, *Beliefs* and *Power* were the two dimensions with significant and relatively large regression coefficients. While *Capacity* had a significant (but weaker) regression coefficient in one of the two model specifications that we tested, *Communion* was never significant. Moreover, *Beliefs* and *Capacity* were the two dimensions with the largest variation between origin groups (Table A9: standard deviations; Table A10: ICCs). Finally, *Power* was the dimensions that correlated least with the other three dimensions (Table 11), which suggests that *Power* stereotypes are important because they add something new to the discussion on how a group is perceived. In addition, *Beliefs* and *Power* were the two content dimensions that emerged in separate factor analyses for most origin groups (Table 10).

3) Stereotypes about immigrant groups vary primarily between groups originating from the poor (and/or Muslim) global south and the wealthy global north: Unfortunately, we found little evidence for ambivalent stereotypes. We could not confirm that immigrant groups that are negatively evaluated on one dimension tend to be positively evaluated on other dimensions (Lee & Fiske, 2006). To the contrary, we observed a stark divide between immigrants from the global south and immigrants from the global north, with the former receiving negative stereotypes and the latter positive stereotypes. Only the stereotypes about Germans (the in-group), Chinese immigrants, and immigrants from the US slightly deviated from this trend: these groups scored high on *Capacity, Beliefs*, and *Power* but only medium on *Communion* (for similar results see Froehlich & Schulte, 2019). This pattern is well documented for Germans and Eastern Asians and it is often interpreted as an example of an ambivalent stereotype ("competent but cold"). However, according to our results, this pattern is solely a variation in the hierarchy *within* the global north; but it does not affect the strong north-south divide.

Summary and Conclusion

This report describes the design and the main results of the *Judging Without Knowing* survey. This survey was composed of two parts that served different purposes. Part one was a photo survey that served as a post-hoc test of the photos that were used as stimulus material in the ADIS and GEMM studies. Part two was a stereotype survey that explored the content of the stereotypes Germans have about Germans and about various immigrant groups in Germany.

The photos survey revealed significant differences between photos with respect to sympathy, attractiveness, and competence. Importantly, however, while there were important differences between photos of males and females and between the photo series (i.e. adjusted ADIS, original ADIS, and GEMM), there were only marginal difference between phenotype groups (i.e. Asian, Black, Southern White, and Northern White). The only exceptions were the attractiveness ratings of the Asian photos: Asian photos received significantly more negative attractiveness ratings. This is, of course, not ideal, because the photos were used as phenotype signals in the ADIS and GEMM studies and were chosen because of their supposed comparability. Fortunately, however, the field experimental results for (Eastern) Asian job applicants were generally quite positive (i.e. a medium to high likelihood of receiving a positive response), which suggests that there were no serious negative biases in consequences the lower attractiveness of "Asian" photos. In sum, the survey ensured that the photos are well-suited as stimulus material in the ADIS and GEMM studies.

The stereotype survey, by contrast, did not test the material that was used in previous studies, but explored and added important knowledge about a potentially relevant factor that might affect ethnic hierarchies in hiring: ethnic stereotypes. To this end, we asked respondents to rate Germans and immigrant groups from 37 different countries of origin on a range of bipolar adjectives that are part of different stereotype content models (SCM: Cuddy et al., 2008; Fiske et al., 2002; facets model: Abele et al., 2016; ABC model: Koch et al., 2016). We found that instructions matter: respondents generally expressed more positive stereotypes when being asked to provide their own opinion but more negative views when being asked to indicate what "Germans think". Second, the four stereotype content dimensions that emerged were only partly reconcilable with the three different stereotype content models, while in some respects they were contradicting of all three stereotype content models. While *Communion* did not appear to be the primary dimension in stereotypes about immigrant groups in

Germany, progressive *Beliefs* and *Power* seem to be of high importance. *Capacity* also received supportive evidence. Finally, the pattern of results revealed a clear divide between immigrants from the global north and immigrants from the global south. Germans and immigrants from the global north were rather positively viewed on all four stereotype content dimensions, while immigrants from the global south were negatively viewed on all four stereotype content dimensions: as rather traditional, powerless, incapable, and cold.

Literature

Abele, A. E., Hauke, N., Peters, K., Louvet, E., Szymkow, A., & Duan, Y. (2016). Facets of the Fundamental Content Dimensions: Agency with Competence and Assertiveness—Communion with Warmth and Morality. *Frontiers in Psychology*, *7*(1810). doi: 10.3389/fpsyg.2016.01810

Agerström, J., & Rooth, D.-O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology*, *96*(4), 790–805. doi: 10.1037/a0021594

Brigham, J. C. (1972). Racial Stereotypes: Measurement Variables and the Stereotype-Attitude Relationship. Journal of Applied Social Psychology, 2(1), 63–76. doi: 10.1111/j.1559-1816.1972.tb01264.x

Burgess, D., & Borgida, E. (1999). Who women are, who women should be: Descriptive and prescriptive gender stereotyping in sex discrimination. *Psychology, Public Policy, and Law*, *5*(3), 665–692. doi: 10.1037/1076-8971.5.3.665

Castilla, E. J. (2008). Gender, race, and meritocracy in organizational careers. *American Journal of Sociology*, 113(6), 1479–526. doi: 10.1086/588738

Cuddy, A. J., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of personality and social psychology*, *92*(4), 631-648. doi: 10.1037/0022-3514.92.4.631

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in experimental social psychology*, *40*, 61–149. doi: 10.1016/S0065-2601(07)00002-0

Devine, P. G., & Elliot, A. J. (1995). Are Racial Stereotypes Really Fading? The Princeton Trilogy Revisited. *Personality and Social Psychology Bulletin*, *21*(11), 1139–1150. doi: 10.1177/01461672952111002

Dovidio, J. F., Evans, N., & Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology*, *22*(1), 22–37. https://doi.org/10.1016/0022-1031(86)90039-9

Fiske, S. T. (2018). Stereotype Content: Warmth and Competence Endure. *Current Directions in Psychological Science*, *27*(2), 67–73. https://doi.org/10.1177/0963721417738825

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

Froehlich, L., & Schulte, I. (2019). Warmth and competence stereotypes about immigrant groups in Germany. *PLOS ONE*, *14*(9), e0223103. doi: 10.1371/journal.pone. 0223103

Gaddis, S. M. (2018). *Audit studies: Behind the scenes with theory, method, and nuance* (Bd. 14). Cham: Springer. doi: 10.1007/978-3-319-71153-9

Glick, P., & Fiske, S. T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American psychologist*, *56*(2), 109-118. doi: 10.1037/0003-066X.56.2.109

Huang, F. L., & Cornell, D. G. (2016). Using Multilevel Factor Analysis With Clustered Data: Investigating the Factor Structure of the Positive Values Scale. *Journal of Psychoeducational Assessment*, *34*(1), 3–14. doi: 10.1177/0734282915570278

Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. British journal of Social Psychology, 33(1), 1–27. doi: 10.1111/j.2044-8309.1994.tb01008.x

Jost, J. T., Kivetz, Y., Rubini, M., Guermandi, G., & Mosso, C. (2005). System-justifying functions of complementary regional and ethnic stereotypes: Cross-national evidence. Social justice research, 18(3), 305–333. doi: 10.1007/s11211-005-6827-z

Kay, A. C., & Jost, J. T. (2003). Complementary Justice: Effects of „Poor but Happy" and „Poor but Honest" Stereotype Exemplars on System Justification and Implicit Activation of the Justice Motive. Journal of Personality and Social Psychology, 85(5), 823–837. doi: 10.1037/0022-3514.85.5.823

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

Kotzur, P. F., Friehs, M.-T., Asbrock, F., & Zalk, M. H. W. van. (2019). Stereotype content of refugee subgroups in Germany. European Journal of Social Psychology, 49(7), 1344–1358. doi: 10.1002/ejsp.2585

Kotzur, P. F., Veit, S., Namyslo, A., Holthausen, M.-A., Wagner, U., & Yemane, R. (2020). „Society thinks they are cold and/or incompetent, but I do not": Stereotype content ratings depend on instructions and the social group's location in the stereotype content space. *British Journal of Social Psychology*. doi: 10.1111/bjso.12375

Lancee, B., Birkelund, G. E., Coenders, M., Di Stasio, V., Fernández Reino, M., Heath, A., Koopmans, R., Larsen, E., Polavieja, J., Ramos, M., Thijssen, L., Veit, S., Yemane, R., & Zwier, D. (2019). *The GEMM study: A cross–national harmonized field experiment on labour market discrimination: Technical report*. doi: 10.2139/ssrn.3398191. Available at SSRN: https://ssrn.com/abstract=3398191

Lee, T. L., & Fiske, S. T. (2006). Not an outgroup, not yet an ingroup: Immigrants in the stereotype content model. *International Journal of Intercultural Relations*, *30*(6), 751–768. doi: 10.1016/j.ijintrel.2006.06.005

Madon, S., Guyll, M., Aboufadel, K., Montiel, E., Smith, A., Palumbo, P., & Jussim, L. (2001). Ethnic and national stereotypes: The Princeton trilogy revisited and revised. *Personality and Social Psychology Bulletin*, *27*(8), 996-1010. doi: 10.1177/0146167201278007

Neumark, D. (2012). Detecting Discrimination in Audit and Correspondence Studies. *Journal of Human Resources, 47*(4), 1128–1157. doi: 10.3368/jhr.47.4.1128

Pager, D. (2007). The use of field experiments for studies of employment discrimination: Contributions, critiques, and directions for the future. *The Annals of the American Academy of Political and Social Science, 609*(1), 104–133. doi: 10.1177/0002716206294796

Painter, M. A., Holmes, M. D., and Bateman, J. (2016). Skin tone, race/ethnicity, and wealth inequality among new immigrants. *Social Forces, 94*(3), 1153–1186. doi: 10.1093/sf/sov094

Phalet, K., & Poppe, E. (1997). Competence and morality dimensions of national and ethnic stereotypes: A study in six eastern-European countries. *European Journal of Social Psychology, 27*(6), 703-723.

Samples S.T. (2019). Black Is *Not* Beautiful: The German Myth of Race. In: Essed P., Farquharson K., Pillay K., White E. (Eds). *Relating Worlds of Racism* (pp. 223-244). Cham: Palgrave Macmillan.

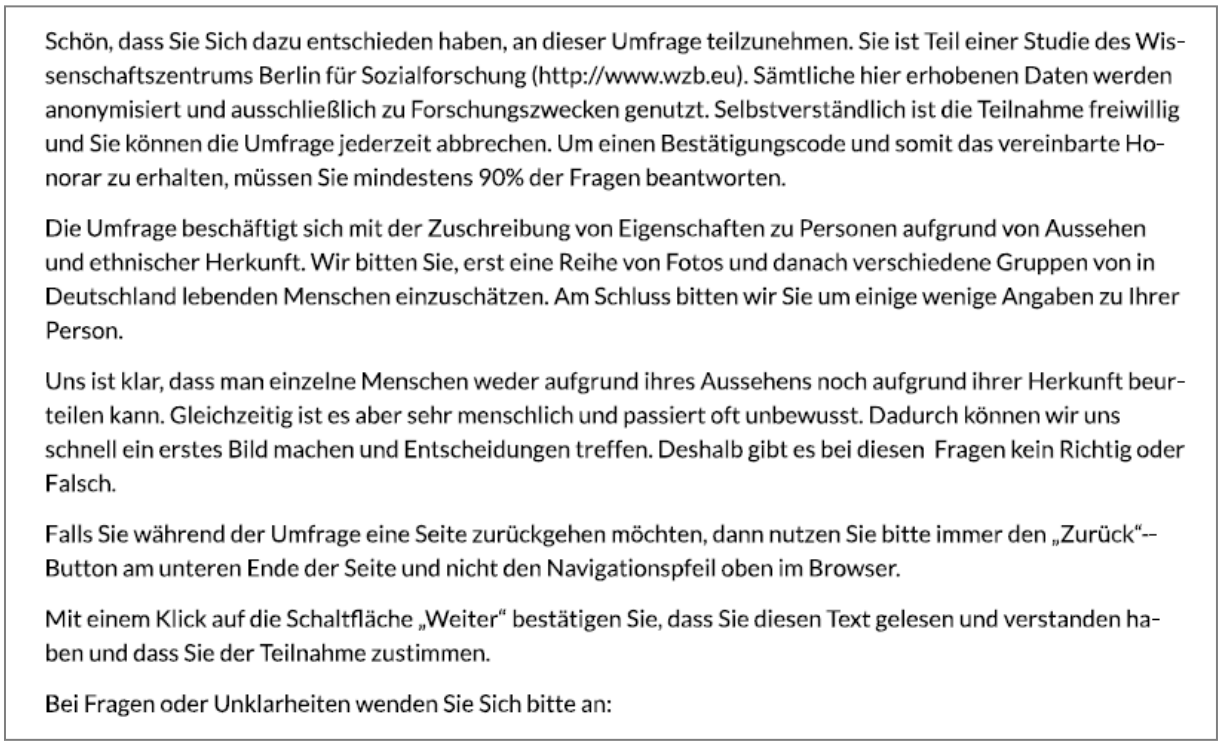
Stangor, C., & Lange, J. E. (1994). Mental Representations of Social Groups: Advances in Understanding Stereotypes and Stereotyping. In M. P. Zanna (Hrsg.), *Advances in Experimental Social Psychology, 26*, 357–416. Academic Press. doi: 10.1016/S0065-2601(08)60157-4

Uhlmann, E., Dasgupta, N., Elgueta, A., Greenwald, A. G., and Swanson, J. (2002). Subgroup Prejudice Based on Skin Color Among Hispanics in the United States and Latin America. *Social Cognition, 20* (3), 198–226. doi: 10.1521/soco.20.3.198.21104

Veit, S., & Yemane, R. (2018). The ADIS study: A large-scale correspondence test on labor market discrimination in Germany (Technical Report), *WZB Discussion Paper VI* 2018-10. Berlin: WZB.

Appendix

Appendix Figures



Schön, dass Sie Sich dazu entschieden haben, an dieser Umfrage teilzunehmen. Sie ist Teil einer Studie des Wissenschaftszentrums Berlin für Sozialforschung (http://www.wzb.eu). Sämtliche hier erhobenen Daten werden anonymisiert und ausschließlich zu Forschungszwecken genutzt. Selbstverständlich ist die Teilnahme freiwillig und Sie können die Umfrage jederzeit abbrechen. Um einen Bestätigungscode und somit das vereinbarte Honorar zu erhalten, müssen Sie mindestens 90% der Fragen beantworten.

Die Umfrage beschäftigt sich mit der Zuschreibung von Eigenschaften zu Personen aufgrund von Aussehen und ethnischer Herkunft. Wir bitten Sie, erst eine Reihe von Fotos und danach verschiedene Gruppen von in Deutschland lebenden Menschen einzuschätzen. Am Schluss bitten wir Sie um einige wenige Angaben zu Ihrer Person.

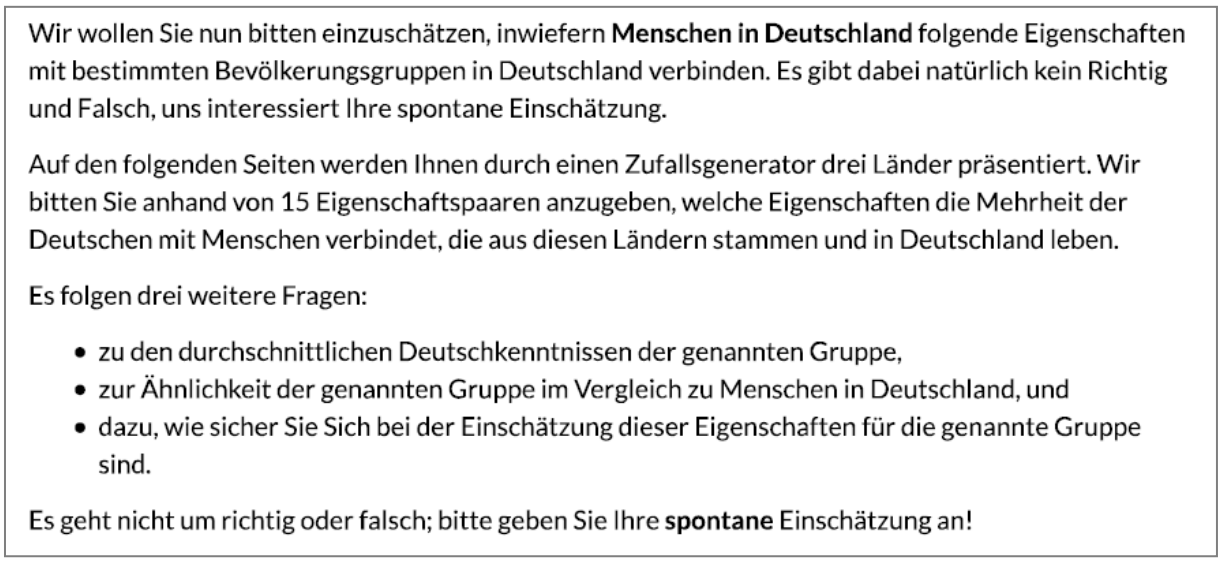
Uns ist klar, dass man einzelne Menschen weder aufgrund ihres Aussehens noch aufgrund ihrer Herkunft beurteilen kann. Gleichzeitig ist es aber sehr menschlich und passiert oft unbewusst. Dadurch können wir uns schnell ein erstes Bild machen und Entscheidungen treffen. Deshalb gibt es bei diesen Fragen kein Richtig oder Falsch.

Falls Sie während der Umfrage eine Seite zurückgehen möchten, dann nutzen Sie bitte immer den „Zurück"-Button am unteren Ende der Seite und nicht den Navigationspfeil oben im Browser.

Mit einem Klick auf die Schaltfläche „Weiter" bestätigen Sie, dass Sie diesen Text gelesen und verstanden haben und dass Sie der Teilnahme zustimmen.

Bei Fragen oder Unklarheiten wenden Sie Sich bitte an:

Figure A1: Screenshot of instruction screen



Wir wollen Sie nun bitten einzuschätzen, inwiefern **Menschen in Deutschland** folgende Eigenschaften mit bestimmten Bevölkerungsgruppen in Deutschland verbinden. Es gibt dabei natürlich kein Richtig und Falsch, uns interessiert Ihre spontane Einschätzung.

Auf den folgenden Seiten werden Ihnen durch einen Zufallsgenerator drei Länder präsentiert. Wir bitten Sie anhand von 15 Eigenschaftspaaren anzugeben, welche Eigenschaften die Mehrheit der Deutschen mit Menschen verbindet, die aus diesen Ländern stammen und in Deutschland leben.

Es folgen drei weitere Fragen:

- zu den durchschnittlichen Deutschkenntnissen der genannten Gruppe,
- zur Ähnlichkeit der genannten Gruppe im Vergleich zu Menschen in Deutschland, und
- dazu, wie sicher Sie Sich bei der Einschätzung dieser Eigenschaften für die genannte Gruppe sind.

Es geht nicht um richtig oder falsch; bitte geben Sie Ihre **spontane** Einschätzung an!

Figure A2: Screenshot of instruction screen – stereotype survey

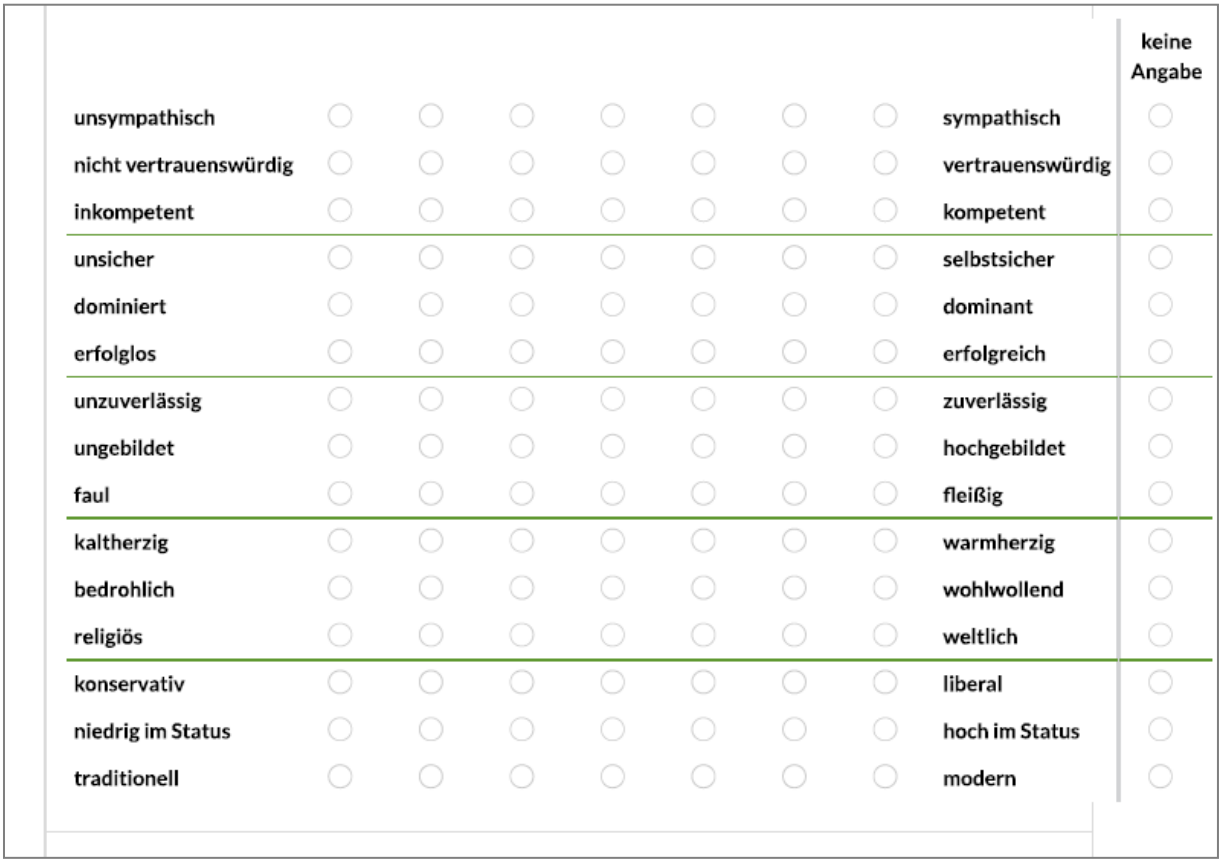


Figure A3: Screenshot of semantic differentials with adjective pairs

Appendix Tables

Table A1: Regression of sympathy ratings for single photos

	(1)	(2)	(3)	(4)	(5)	(6)
	Sympathy original ADIS female	**Sympathy adjusted ADIS female**	**Sympathy GEMM female**	**Sympathy original ADIS male**	**Sympathy adjusted ADIS male**	**Sympathy GEMM male**
Phenotype (ref: Central European)						
Northern Europe	.005	−.058	−.011	−.024	.161***	.137***
	(.113)	(.128)	(.121)	(.133)	(.130)	(.132)
Southern Europe	.014	−.078*	−.093*	−.006	.119***	.207***
	(.118)	(.119)	(.124)	(.130)	(.129)	(.131)
North African Turkish			−.046			.280***
			(.122)			(.131)
East African	−.086***	.058	.027	.004	.174***	.210***
	(.115)	(.110)	(.139)	(.132)	(.130)	(.133)
West African	−.025	.023	−.004	.035	.170***	.248***
	(.118)	(.111)	(.144)	(.132)	(.130)	(.130)
East Asian	−.110***	−.036	−.058	.026	.251***	.247***
	(.112)	(.113)	(.140)	(.129)	(.131)	(.131)
South-East Asian	−.144***	−.223***	−.021	−.053	.112***	.104***
	(.108)	(.116)	(.126)	(.130)	(.124)	(.145)
N_{obs}	2287	1812	1961	1631	1713	1911
R^2	.053	.095	.035	.035	.062	.067

Standardized beta coefficients; Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A2: Regression of attractiveness ratings for single photos

	(1)	(2)	(3)	(4)	(5)	(6)
	Attractive_ness original ADIS female	**Attractive_ness adjusted ADIS female**	**Attractive_ness GEMM female**	**Attractive_ness original ADIS male**	**Attractive_ness adjusted ADIS male**	**Attractive_ness GEMM male**
Phenotype (ref: Central European)						
Northern Europe	−.085***	.004	.006	−.105***	.188***	.168***
	(.112)	(.141)	(.127)	(.130)	(.128)	(.138)
Southern Europe	.011	−.043	.042	.019	.078*	.106**
	(.117)	(.132)	(.129)	(.126)	(.127)	(.137)
North African Turkish			−.018			.222***
			(.128)			(.136)
East African	−.106***	.090**	.007	−.091**	.195***	.010
	(.116)	(.122)	(.145)	(.128)	(.129)	(.139)
West African	−.041*	.022	−.054	−.40**	.189***	.073*
	(.117)	(.123)	(.150)	(.129)	(.128)	(.136)
East Asian	−.151***	−.086**	−.136***	−.081*	.175***	.001
	(.112)	(.125)	(.146)	(.125)	(.129)	(.136)
South-East Asian	−.160***	−.121***	−.091**	−.183***	.160***	.059*
	(.108)	(.128)	(.131)	(.126)	(.122)	(.152)
N_{obs}	2281	1800	1958	1620	1699	1887
R^2	.070	.100	.055	.070	.073	.071

Standardized beta coefficients; Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A3: Regression of competence ratings for single photos

	(1)	(2)	(3)	(4)	(5)	(6)
	Sympathy original ADIS female	Sympathy adjusted ADIS female	Sympathy GEMM female	Sympathy original ADIS male	Sympathy adjusted ADIS male	Sympathy GEMM male
Photo (ref: Central European)						
Northern Europe	-.071***	-.048	.112**	.010	.039	.143***
	(.108)	(.130)	(.116)	(.121)	(.120)	(.128)
Southern Europe	-.005	.106**	.093*	-.031	-.052	.129***
	(.114)	(.122)	(.118)	(.118)	(.120)	(.126)
North African Turkish			.080*			.167***
			(.117)			(.126)
East African	-.110***	.158***	.051	-.007	.052	.064*
	(.113)	(.112)	(.133)	(.119)	(.120)	(.129)
West African	-.045*	.133***	-.005	.008	.052	.065
	(.114)	(.113)	(.137)	(.120)	(.121)	(.126)
East Asian	-.096***	.124***	.048	-.041	.120***	.195***
	(.110)	(.116)	(.134)	(.118)	(.121)	(.126)
South-East Asian	-.100***	.072*	.104**	-.112**	.058	.087**
	(.104)	(.119)	(.119)	(.117)	(.114)	(.141)
N_{obs}	2,220	1,766	1,922	1,596	1,674	1849
R^2	0.045	0.062	0.045	0.025	0.028	0.040

Standardized beta coefficients; Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A4: Regression with covariates

		(1) Sympathy all	(2) Attractiveness all	(3) Competence all
PHOTOS				
Phenotype (ref: Northern White))				
	Asian	-.031***	-.092***	.007
		(.0288)	(.0300)	(.0265)
	Black	.064***	.003	.017
		(.0365)	(.0380)	(.0335)
	Southern White	.044***	.021*	.014
		(.0323)	(.0337)	(.0297)
Photo series (vs. ADIS adjusted)				
	ADIS original	.113***	.090***	.089***
		(.0274)	(.0286)	(.0251)
	GEMM	.168***	.096***	.147***
		(.0273)	(.0284)	(.0249)
Gender on photo (vs. female)				
	Male	-.157***	-.172***	-.128***
		(.0221)	(.0230)	(.0202)
COVARIATES				
Age in years		.097***	.152***	.027
		(.00221)	(.00228)	(.00221)
Origin (ref. migrant):	native	.031	.019	.046**
		(.0601)	(.0620)	(.0604)
Gender (ref. female):	Male	-.069***	-.037*	-.096***
		(.0448)	(.0462)	(.0449)
Other		-.047**	-.040*	-.043*
		(.557)	(.568)	(.548)
Education (ref: low)				
	Higher entrance qualification	.000	-.016	-.013
		(.0834)	(.0861)	(.0840)
	BA or vocational training	.015	-.004	.004
		(.0746)	(.0770)	(.0749)
	MA or higher	-.008	-.011	-.039
		(.0759)	(.0785)	(.0764)
N_{obs}		9,898	9,835	9,652

Standardized beta coefficients; Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A5: Interaction phenotype-by-gender

	(1) Sympathy	(2) Attractiveness	(3) Competence
Phenotype (ref: Northern White))			
Asian	-.0582	-.230***	.0137
	(.0401)	(.0417)	(.0369)
Black	.410***	.205***	.154**
	(.0514)	(.0535)	(.0471)
Southern White	.293***	.260***	.183***
	(.0449)	(.0467)	(.0411)
COVARIATE			
Male respondent (vs. female)	-.0755	.0504	-.180***
	(.0530)	(.0548)	(.0519)
INTERACTIONS			
Asian * male respondent	-.0743	-.123*	.0194
	(.0558)	(.0580)	(.0511)
Black * male respondent	-.295***	-.377***	-.172**
	(.0712)	(.0741)	(.0652)
Southern White * male respondent	-.280***	-.367***	-.272***
	(.0620)	(.0644)	(.0567)
N_{obs}	9894	9830	9647
R^2	.074	.074	.042

Regression coefficients; Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A6: Interaction gender-by-gender

	(1) Sympathy	(2) Attractiveness	(3) Competence
Male person on photo (vs. female)	-.437***	-.572***	-.417***
	(.0311)	(.0323)	(.0284)
COVARIATE			
Male respondent (vs. female)	-.192***	-.177***	-.327***
	(.0490)	(.0506)	(.0484)
INTERACTIONS			
Male * male respondent	-.003	.148**	.162***
	(.044)	(.045)	(.040)
N_{obs}	9894	9830	9647
R^2	.064	.071	.041

Regression coefficients; Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A7: Multilevel factor analyses with 15 descriptive adjectives (MEFA)

	Within		**Between**		**X2**	**df**	**p**	**RMSEA**	**CFI**	**TLI**	**SRMRw**	**SRMRb**	**AIC**	**BIC**
	N	Eigen-value	N	Eigen-value										
MEFA	1	**6.85**	1	**11.17**	7398.429	180	0.0000	0.089	0.816	0.785	**0.078**	0.134	219808.290	220297.409
	2	**1.50**	1		4850.056	166	0.0000	0.075	0.880	0.849	**0.054**	0.133	217287.917	217868.338
	3	**1.30**	1		2140.846	153	0.0000	**0.051**	**0.949**	0.930	**0.032**	0.128	214604.707	215269.908
	4	.93	1		1294.793	141	0.0000	**0.040**	**0.971**	**0.956**	**0.013**	0.128	213782.654	214526.115
	5	.57	1		952.679	139	0.0000	**0.035**	**0.979**	**0.966**	**0.008**	0.128	213462.540	214277.738
	1-5		2	**1.68**	no convergence									
	1-5		3	**1.45**	no convergence									
	1		4	.52	6821.384	141	0.0000	0.097	0.829	0.746	**0.078**	**0.006**	219309.245	220052.705
	2		4		274.451	127	0.0000	0.081	0.894	0.825	**0.054**	**0.006**	216790.312	217625.074
	3		4		1575.339	114	0.0000	**0.051**	**0.963**	0.931	**0.032**	**0.005**	214117.200	215036.743
	4		4		730.666	102	0.0000	**0.035**	**0.984**	**0.967**	**0.013**	**0.005**	213296.527	214294.329
	5		4		387.571	91	0.0000	**0.025**	**0.992**	**0.983**	**0.008**	**0.005**	212975.432	214044.972
	1-5		5	.06	no convergence									

Cut-off criteria for good fit x2 < .05; RMSEA < .08; CFI >=.90; TLI >=.95; SMRM <.08 (see e.g. https://www.cscu.cornell.edu/news/Handouts/SEM_fit.pdf).
Fit indices in bold meet the cuff-of criteria.

Table A8: MEFA factor loadings: 5 within and 4 between factors

		Within					Between			
		1: Power	*2: Beliefs*	*3: Status*	*4: Commu-nion*	*5: Capa-city*	*1: Power*	*2: Capacity*	*3: Beliefs*	*4: Commu-nion*
Items	Factor loadings									
1	Competent					.490		.982		
2	Laborious					.799		1.095		
3	Reliable					.780		.936		
4	Educated			.624				.990		
5	High in Status			.764				.771		
6	Successful			.546				.963		
7	Modern		.769						.679	
8	Secular		.641						.876	
9	Liberal		.646						.756	
10	Dominant	.507					1.028			
11	Self-confident	.774					.802			
12	Warm				.775					1.138
13	Trustworthy				.448			.720		
14	Likeable				.759					.650
15	Benevolent				.575					.501

Note: Factor loadings smaller than .40 are not shown.
For each item, the highest factor lading is highlighted in bold.
Items with substantial cross-loading are highlighted in italics.

Table A9: Stereotype content dimensions by origin groups

Groups' mean	N_{groups}	Mean	Std. Dev.	Min	Max
Capacity	38	4.25	.81	3.12	5.91
Power	38	4.39	.51	3.59	5.64
Beliefs	38	3.45	.83	2.27	5.48
Communion	38	4.52	.56	3.55	5.49

Table A10: Empty models

DV: Stereotype content dimension	$N_{observations}$	N_{groups}	ICC
Capacity	4,999	38	.38
Power	4,946	38	.16
Beliefs	5,002	38	.31
Communion	4,988	38	.19