In your face

A comparative field experiment on racial discrimination in Europe

Polavieja, J.G.; Lancee, B.; Ramos, M.; Veit, S.; Yemane, R.

DOI
10.1093/ser/mwad009

Publication date
2023

Document Version
Final published version

Published in
Socio-Economic Review

Socio-Economic Review, 2023, Vol. 21, No. 3, 1551–1578
https://doi.org/10.1093/ser/mwad009
Advance Access Publication Date: 25 March 2023
Article

Article

In your face: a comparative field experiment on racial discrimination in Europe

Javier G. Polavieja ,[1,*] Bram Lancee ,[2] María Ramos ,[3] Susanne Veit ,[4] and Ruta Yemane [4]

[1]Universidad Carlos III de Madrid, Department of Social Sciences, Getafe, Spain, [2]University of Amsterdam, Department of Sociology, The Netherlands, [3]Department of Analysis and Research, Office of the Prime Minister of Spain, Madrid, Spain, [4]DeZIM-Institute, Berlin, Germany

**Correspondence: javier.polavieja@uc3m.es*

Abstract

We present the first large-scale comparative field experiment on appearance-based racial discrimination in hiring conducted in Europe. Using a harmonized methodology, we sent fictitious résumés to real vacancies in Germany, the Netherlands and Spain, randomly varying applicants' ethnic ancestry (signaled foremost by name) and applicants' racial appearance (signaled by photographs). Applicants are young-adult country nationals born to parents from over 40 different countries of ancestry ($N = 12\,783$). We examine average differences in callback across four phenotypic groups and four regions of ancestry and present the first cross-country comparable estimates of appearance-based racial discrimination reported in the field-experimental literature. We find that applicants' phenotype has a significant and independent effect on employers' responses in Germany and the Netherlands, whereas in Spain we only find evidence of hiring discrimination for particular combinations of phenotype and ancestry, which suggests a less direct and more complex effect of phenotype in this country. Implications are discussed.

Keywords: race, ethnicity, discrimination, employer behavior, Europe, field experiments

1. Introduction

Because of increasing migration inflows, Europeans today are phenotypically more diverse than ever in modern history and the number of Europeans potentially at risk of experiencing appearance-based racial discrimination is on the rise.[1] There are over 19 million Europeans

1 Throughout this study, we use the terms 'phenotype' and 'racial appearance', and the related terms 'phenotypic discrimination' and 'appearance-based racial discrimination' interchangeably.

of foreign-born parents in the European Union (5 million of whom are of non-European ancestry) and almost half of them (9 million) are employed (EUROSTAT, 2017). Studying appearance-based racial discrimination in employment is thus crucial for understanding the barriers that Europeans of foreign ancestry are likely to face and, more generally, the mechanisms of social stratification currently operating in ethnically diverse European societies.

However, most research on racial discrimination has been conducted in the USA, where the racial divide has historically shaped the majority–minority group relations and remains up until today the dominant social cleavage (Rumbaut and Portes, 2001; Alba, 2005). This is in stark contrast to (continental) Europe, where 'race' has largely been an 'anathema to official thinking' (Thomson and Crul, 2007). Research on immigrants' incorporation in Europe has identified religion as the key ethnic boundary (Alba, 2005) and, more specifically, a Muslim background as a key driver of discrimination (Strabac and Listhaug, 2008; Pierné, 2013; Wimmer and Soehl, 2014; Adida *et al.*, 2016; Valfort, 2017; Di Stasio *et al.*, 2021). The academic literature has largely ignored the role of phenotype and the extent to which being a 'visible' minority is an additional source of discrimination. Yet, evidence shows that immigrants and their descendants do feel discriminated against in Europe because of their 'race' (see, e.g. Connor and Koenig, 2015; Flores, 2015). Despite early warnings that a 'new racism' could be emerging across European societies (see, e.g. Balibar, 1991), to date, the empirical study of appearance-based racial discrimination has been widely neglected by continental European scholars,[2] a neglect that mirrors what Goldberg (2006) called the 'official denial of race'.

To fill this gap, we ran the first large-scale comparative field experiment on appearance-based racial discrimination in hiring conducted in Europe. The experiment was carried out simultaneously and with a fully standardized design in three European countries, Germany, the Netherlands and Spain. Using an unpaired design (i.e. one job application per company), we sent almost 13 000 fictitious résumés to real vacancies, randomly varying applicants' ethnic ancestry (signaled using ethnic-sounding names as well as information on parents' country of origin and mother tongue) and phenotype (signaled using applicants' photographs). A unique contribution of this study is that we can test for the causal effect of racial appearance (phenotype) and ancestry as potentially distinct triggers of hiring discrimination. More specifically, we investigate whether European employers discriminate applicants of particular phenotypes and whether this discrimination operates independently or in interaction with applicants' ancestry. This, we claim, is an important contribution to the discrimination literature for the following two reasons: *First*, because existing field-experimental research on racial discrimination has typically used applicant's name as the only signal of 'race', thus confounding ancestry and phenotypical triggers of discrimination in one single treatment, as we explain below. *Second*, because, in the specific European context, the neglect of 'race' in

Phenotype (racial appearance) refers to 'aspects of a person's physical appearance that are socially understood as relevant to racial classification' (Roth, 2016, p. 1317). This includes skin tone as well as other features such as hair texture and color, nose and lips shape and eye color and shape. Monk (2016, p. 414) offers a similar definition for 'color' as a concept 'that includes skin color and other ethnoracially coded phenotypical traits'.

2 British scholarship stands out from the rest of Europe in that it has traditionally paid greater attention to racial discrimination (for a review, see, e.g. Riach and Rich, 2002). We note, however, that researchers cannot draw on correspondence tests to measure the distinctive role of applicant's appearance in the UK because British résumés do not include photographs.

empirical research could have led to overestimating the role of culture and religion as triggers of prejudice and discrimination. In other words, the argument that discrimination against non-European descendants in Europe has mostly cultural–religious basis implicitly assumes that Europeans are mostly 'color blind'. We test this assumption empirically and prove it wrong. A final contribution of this study is its innovative comparative dimension, which allows us to investigate whether there are significant differences across the three countries studied, thus contributing to the expanding body of literature on hiring discrimination and the role of 'race' outside the USA.

2. Framework

Research on racial discrimination has drawn heavily on the US case. This is hardly surprising given the central role that race has historically played as a determinant of Americans' life-chances. Racial discrimination in the USA has been extensively investigated in sociology (see reviews in, e.g. Pager and Shepherd, 2008; Brubaker, 2009; Reskin, 2012), economics (see reviews in, e.g. Charles and Guryan, 2011; Bertrand and Duflo, 2017), social psychology (see reviews in, e.g. Mays *et al.*, 2007; Richeson and Sommers, 2016) and cognitive psychology (see reviews in, e.g. Macrae and Bodenhausen, 2000; Phelps and Thomas, 2003; Maddox, 2004).

Roth (2016) discusses the multiple dimensions of the concept 'race', as typically used in the US context, including racial identity, racial ancestry, observed race and phenotype, among others, and stresses that different dimensions of race can yield different estimates of racial inequality. Two dimensions discussed by Roth are particularly relevant to the study of racial discrimination: observed race (i.e. the race observers perceive others to be) and phenotype (the physical traits observers use as the main clues for racially classifying others). Phenotype—or 'color' (Monk, 2016, p. 414)—is therefore the main determinant of observed race. Interestingly, however, correspondence tests on employment discrimination conducted in the USA cannot manipulate applicants' phenotype because résumés do not include photographs in this country. Instead, researchers signal race using 'racially' typical names (see, e.g. Bertrand and Mullainathan, 2004). This means 'observed' race cannot be observed at all—but only *imagined*—by US employers. The use of racially sounding names as the only signal of 'race' inevitably confounds ancestry and (imagined) phenotype (and also likely socioeconomic background signals) in one single treatment (see Gaddis, 2017).

Most research on the impact of phenotype (racial appearance) in the USA has thus been conducted within the extensive observational literature on colorism, which primarily focuses on how skin tone differences can affect lifetime prospects even within members of the same self-identified racial group (for a review see Dixon and Telles, 2017). A large body of evidence based on survey data (especially the National Survey of Black Americans) suggests that darker skin tones (externally measured by interviewers using skin-tone scales) are negatively correlated with measures of educational achievement, socioeconomic status and spousal status among respondents who self-define as African Americans (see, e.g. Gullickson, 2005; Goldsmith *et al.*, 2006; Hersch, 2006; Monk, 2014). These findings are consistent with social-psychological laboratory research on the so-called phenotypic prototypicality (PP), which shows that African American targets with high PP (including darker skin) are described by perceivers using more stereotypic traits than low PP African Americans (see Maddox, 2004; Maddox and Perry, 2018).

An important limitation of the existing research on colorism, however, is that surveys seldom include full information on respondents' skills and socioeconomic family background. This might lead to confounding racial discrimination processes with well-known processes of intergenerational class reproduction—which could perpetuate socioeconomic disparities across people of different racial appearances even in the absence of racial discrimination (see, e.g. Flores and Telles, 2012; Wimmer, 2015). Another important limitation is the subjective nature of color measures and, in particular, the possibility that target's socioeconomic status taints observers' perceptions of color (see, e.g. Freeman *et al.*, 2011; Flores and Telles, 2012; Saperstein and Penner, 2012).

Hence, despite the centrality of race in US scholarship, empirical research on racial discrimination straddles between two methodological limitations: one stemming from the use of survey data to measure discriminatory behavior, the other stemming from the impossibility of manipulating applicants' physical appearance in US résumés. A key methodological advantage of the present study over the preceding literature is that we can manipulate applicants' photographs and applicants' ethnic ancestry *separately*—within the limits of plausibility. This allows us to study the interplay of phenotype and ancestry as triggers of discrimination, as well as to estimate *net effects* of applicants' phenotype on employers' callbacks across three different European countries—to our knowledge for the first time in the field-experimental literature.

2.1 Detecting racial discrimination in Europe

The study of racial discrimination in continental Europe has been seriously hampered by the lack of data. National statistical offices typically do not collect information on people's racial identification, not even on their ethnic ancestry, as this is considered illegal, which reflects the lack of institutionalization of racial categories in mainland Europe (Simon, 2012; Farkas, 2017).[3] Furthermore, European scholars have often questioned the use of race as a category of analysis. For example, in a famous critique, Bourdieu and Wacquant (1999) argued that the application of race-relation models grounded in the US tradition to the study of other societies was but a form of 'cultural imperialism' and 'ethnocentric intrusion'. In a less passionate vein, Wimmer (2015) provides a contemporary critique of what he calls 'the paradigm of race-centrism', which he considers dominant in the US literature (for similar critiques, see Loveman, 1999; Brubaker, 2009; Banton, 2012; Wimmer, 2013). A core point of critique specifically concerns the role of race outside the USA. Because the USA is an extreme 'racialized social system' (Bonilla-Silva, 1997), these authors argue, the shape and intensity of racial categorization, as well as its consequences for people's life-chances and social identities, are unlikely to travel to other less extreme social contexts (see Loveman, 1999; Wimmer, 2013).

3 Information on racial identity is universally absent from surveys across Europe with the only exception of the UK. To our knowledge, no survey has ever included information on observed race (i.e. phenotypic traits such as skin color measured by the interviewer) in either continental Europe or the UK. To be sure, there are strong normative reasons to reject the institutionalization of racial categories, for such institutionalization could itself act as a powerful source of legitimation of racial ideologies. The horrors of European modern history attest to the perils of institutionalized racial thought—to which European scholarship contributed so decisively in the past (Winant, 2017). The denial of race in Europe must therefore be understood in the historical context.

In contrast to this view stressing US exceptionalism, the so-called race formation scholars contend that both racial ideologies, as well as the racial hierarchies these ideologies help sustain, are universally shared because they are rooted in Europeans' colonization of the Global South, a truly global historical process in which the slave trade played a key role (see, e.g. Winant, 2017). This historical process, the argument goes, would have spread racial categories across the globe, placing 'White' and the top and 'Black' at the bottom of a universally shared racial ranking. Racial-formation theories thus see race as a powerful factor of stratification, capable of producing the starkest forms of social exclusion, not only in the USA, but also around the world. While the last two decades have witnessed an increase in (survey-based) research on racial stratification in Latin American countries (see Telles and the Project on Ethnicity and Race in Latin America (PERLA), 2014; Dixon and Telles, 2017), to date, theoretical debates on the role of race as a source of inequality and discrimination remain largely untested in continental Europe.[4]

To our knowledge, there are only three published studies specifically addressing phenotypic discrimination in hiring in mainland European countries. Weichselbaumer (2017) studied ethno-racial discrimination in hiring in Austria using applicants' photographs in addition to ethnic-sounding names for five different representative ancestries (Austrian, Serbian, Turkish, Chinese and Nigerian). She found significant discrimination against applicants of all non-native ancestries, particularly against Nigerian-ancestry applicants. However, in Weichselbaumer's design, applicants' photographs did not vary within ethnic ancestry and hence the reported discrimination estimates confound phenotype and ancestry. Derous *et al.* (2017) carried out a vignette study on White majority HR professionals in Belgium and found equally qualified applicants with a dark skin tone received lower job suitability ratings than applicants with a light skin tone and this regardless of their ethnic ancestry (signaled using Arab-sounding and Flemish-sounding names). Finally, in the closest experiment to our own, Koopmans *et al.* (2019) investigated the responses of German employers to almost 6000 fictitious German-born applicants from 35 different ancestry countries, 4 religious affiliations and 3 phenotypes (White, Asian and Black). They found significant discrimination against Black applicants (but not against applicants with the Asian phenotype). It must be noted, however, that Koopman *et al.*'s estimates are based on the assumption that applicants' phenotype has identical effects on employers' responses regardless of applicants' ethnic ancestry. We test this assumption empirically—to our knowledge for the first time in the field-experimental literature—by exploiting full phenotypic variation for four phenotypic groups across four large regions of ancestry, as explained below.

2.2 Research questions

The unique research design of this study allows us to overcome the above-mentioned methodological limitations of the US literature and to inaugurate a new research agenda for the

4 In their review on colorism research, Dixon and Telles (2017) note that 'race' (understood as ancestry-based identity) and 'color' (understood as phenotype) might be more easily distinguishable dimensions in the USA, where the one-drop rule (which assigned black status to anyone with even just one African foreparent) institutionalized rigid racial categories. In contrast, in much of Latin America, the word 'race' is hardly used to categorize people and color is the primary basis for ethno-racial categorization, which is comparatively more fluid and has a more clearly gradational nature.

comparative study of racial discrimination using field-experimental methods. We address the following four research questions: (a) to what extent does applicant's racial appearance affect European employers' hiring decisions?; (b) are the effects of racial appearance independent from the effects of applicants' ethnic ancestry or do ancestry and phenotype interact as triggers of discrimination?; (c) how do our appearance-based racial discrimination estimates compare to the existing (confounded) estimates typically reported in US correspondence tests? and (d) are there significant and interpretable differences in the role of racial appearance across the three countries of our study?

Expected country differences

We must recognize from the outset that the answer to the question of expected country differences can only be tentative, as there are more factors potentially affecting variation in racial discrimination than countries in our comparative correspondence test. Yet, this does not imply that we cannot provide some theoretically grounded expectations for country-level variation in discrimination rates. To this end, we focus on the classical theories on inter-group relations, assuming, for now, the historical heritage of colonialism exerts a largely homogeneous cultural influence on all three countries as former European colonial powers.[5]

Drawing on the classical literature on prejudice and inter-group conflict, we should expect higher levels of racial discriminations in contexts where (a) the potential for positive social contact with racial minorities is lower (see, e.g. Allport, 1954) and (b) the degree of competition between majority and minority populations is higher (e.g. Blalock, 1967). Germany and the Netherlands are old immigration countries,[6] and this means German and Dutch employers have likely had greater opportunities for positive social contact with minority populations than Spanish employers. Spain, on the other hand, is a new immigration country, which received a huge inflow of international migrants in a recent and very short period of time. Between the late 1990s until the beginning of the Great Recession, Spain received as many as 5 million international migrants. This made Spain the second most popular destination for international migrants in the world in absolute terms after the USA, leading to a spectacular increase in population diversity (including phenotypic diversity). This unprecedented influx of migrants was immediately followed by one of the most severe economic recessions experienced in Europe after 2008. According to classical group-conflict theories, this combination should provide the perfect storm for the rise of prejudicial attitudes and discrimination against new visible minorities in Spain—as compared to Germany and the Netherlands (see also Polavieja, 2016). Because our fieldwork experiment uses a fully standardized design and because it was conducted simultaneously in all three countries, we can test for these macro-level expectations, as a first explorative take on cross-national variation in racial discrimination in Europe.

5 We revisit this assumption critically in the concluding section.

6 In the European context, 'old' immigration countries are defined as those that showed a considerable and systematic surplus of immigration over emigration (i.e. steady positive net migration) already in the post-World War II period (particularly between 1955 and 1975). Due to this post-war immigration wave, old immigration countries in Europe have sizable minorities of immigrant descent (see Fassmann and Reeger, 2012, p. 66).

3. Data and method

We ran the first cross-national harmonized field experiment on racial discrimination in hiring. The experiment was carried out as part of the GEMM project, a large-scale comparative project on employment discrimination against Europeans of immigrant background (for further details on the design and implementation see Lancee *et al.*, 2019*a*,*b*).[7] In Germany, Spain and the Netherlands, we sent out fictitious cover letters and CVs in response to almost 13 000 real vacancies advertised on online job-search platforms. Following studies such as Ahmed *et al.* (2013), Weichselbaumer (2017) and Koopmans *et al.* (2019), we used an unpaired design and sent one application to each vacancy. The data collection took place over a time span of 2 years (from 2016 to 2018). We targeted a total of 12 783 firms: 3161 in Germany, 5220 in Spain and 4402 in the Netherlands.

In all three countries, we applied to the same seven different occupations, which account roughly for 15–20% of the national workforces. We included five jobs requiring low or middle education levels (ISCO-08 codes in parentheses): Cook (512), hairdresser (5141), payroll clerk (2411, 3313, 411, 412), receptionist (422) and store assistant (522). We also included two occupations requiring education up to a bachelor's degree: sales representative (3322) and software developer (252). The occupations were selected for variation in educational requirements, skills and customer contact, yet the actual distribution of job applications for each occupation varies by country due to differences in the structure of demand.[8] Besides the occupation-specific qualifications, job applicants have a standard profile. All applicants are citizens of the country of study and are between 22 and 26 years old. Yet, their parents come from over 40 different countries of ancestry (including Germany, the Netherlands and Spain). Candidates have obtained the necessary educational qualifications for their occupation in their country (i.e. Germany, the Netherlands or Spain, as applicable) and have 4 years of working experience at two different companies in the same sector of the job vacancy (see Supplementary Box A1 for an example of CV and cover letter). More detailed information on the experimental design can be found in the technical report, which is available online (see Lancee *et al.*, 2019*b*).

3.1 Measurement

Phenotype

Each CV contains a randomly assigned profile picture. For both men and women, we constructed eight pictures that stand for four phenotypic groups, which correspond to 'folk' perceptions of 'racial groups'[9] and which we label: Black, Asian/Indigenous (ASIN), Dark-

7 The full GEMM study includes five European countries: Germany, the Netherlands, Spain, the UK and Norway. Yet, the racial discrimination experiment was conducted only in the former three countries, as only in these countries it is a sufficiently common practice to attach a photograph to applicants' CVs.

8 In the Netherlands, three additional occupations (i.e. electrician, carpenter and plumber) were considered. Models can be tested using fixed effects for occupations, which results in the loss of observations for the Netherlands in pooled models or, alternatively, by using controls for the skill and customer contact requirements of occupations, i.e. using occupational clusters. Results are fully comparable across these two alternative specifications.

9 Racial groups are social constructs people use for categorizing others (National Research Council, 2004). Hence, by using this term we do not wish to imply that these are internally self-aware entities on whose boundaries all actors agree.

Skinned Caucasian (DSC) and White.[10,11] An overview of the pictures is presented in Figure 1. Pictures are standardized in attire and carefully matched in dimensions of attractiveness, competence and sympathy.[12] This makes candidates comparable in appearance dimensions other than 'color'. Based on a pretest, randomization of pictures was restricted to plausible ancestry–phenotype combinations (e.g. a Black phenotype is not assigned to a job applicant of Polish-born parents). Supplementary Table A3 shows the phenotypes that were considered plausible for each ethnic ancestry.[13]

Ethnic ancestry

Ethnic ancestry is measured as the country of origin of (parents of) the job applicant. In total, 44 different ancestries are included. The selection of countries of ancestry in the GEMM study was based on geographic and phenotypic variation as well as on the size of the minority group in each country of study (see Lancee *et al.*, 2019*b*). Within the following strata, ancestry was randomly assigned: 25% majority population (i.e. children of native parents); 25% for two minority groups of great demographic relevance in each country (12.5% for each minority groups of special relevance); 50% for the remaining ancestries. The minority groups of special relevance are Turkish- and Lebanese-ancestry applicants in Germany, Turkish- and Moroccan-ancestry applicants in the Netherlands, and Ecuadorian- and Moroccan-ancestry applicants in Spain. Among the 50% of all applicants from remaining ancestries, we include the same 31 different countries of descent in all three countries of the experiment, plus five additional ancestries chosen to ensure that minority populations of particular interest in the respective national context were included (see Supplementary Table A2 for a list of all countries used). The experiment was thus designed to conciliate two goals: (a) allowing for cross-country comparisons and (b) producing externally valid discrimination estimates for meaningful minorities in each country.

10 'Asian/Indigenous' includes two partly but not completely overlapping phenotypic groups: ASIN 1 and ASIN 2. Photograph validity pre-tests (see footnote 13) suggested phenotype ASIN 1 is plausible for people of East Asian ancestry, while phenotype ASIN 2 has a much broader regional scope, as it is plausible for applicants of East Asian, Southeast Asian, Middle Eastern and Latin American/ Caribbean descent (see Supplementary Table A3). When estimating net phenotypic effects, we group ASIN 1 and ASIN 2 together to maximize statistical power. Yet, we note results are fully robust to restricting the analysis to ASIN 2 alone (available upon request).

11 While the ASIN 2 phenotype might not be the most representative phenotype of particular ethnic groups in particular regions (e.g. Quechua populations in Andean countries), it is a plausible phenotype for all these regions of ancestry (including Andean countries), as shown by our plausibility pretest.

12 To validate the comparability of photographs, a pre-test survey was carried out with a German online sample of 2296 participants. This test was later complemented with a larger online photograph survey with over 5000 participants (see Veit and Yemane, 2020 and Supplemenatry Table A1). Further, photograph-comparability tests replicating the original German survey were carried out in all three countries using smaller convenience samples ($N \approx 50$ in each country) (see Lancee *et al.*, 2019*b*).

13 A phenotype–ancestry plausibility picture pre-test was carried out with a convenience sample of native participants from the five countries of the GEMM study. Over 200 testers indicated for eight photos of either female or male persons whether they consider certain countries of origin as implausible for the person portrayed (see Lancee *et al.*, 2019*b*).

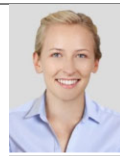
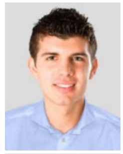
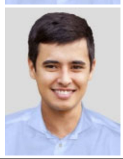
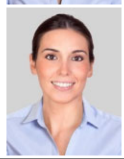
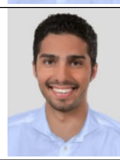
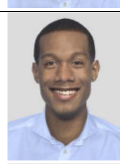
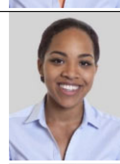
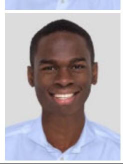
Code	Phenotype	Male	Female
1.1	White 1		
1.2	White 2		
1.3	White 3		
2	DSC		
3.1	ASIN 1		
3.2	ASIN 2 (indigenous)		
4.1	Black 1		
4.2	Black 2		

Figure 1 Photographs used in the experiment.

Information on the country of ancestry is conveyed using three simultaneous signals. First, each minority group was assigned a family and first name typical for the majority population of the ancestry country. Names were chosen that are popular and recognizable as male or female but have no strong religious or class connotation (see Gaddis, 2017).[14] Supplementary Table A2 lists all 'ethnic' names used in this study. Second, the country of ancestry was mentioned explicitly in the cover letter, which contains a statement that the family of the job candidate has migrated from the ancestry country to the region of the advertised job. Third, whenever this was applicable, in the skills section of the CV, the applicant's country of ancestry was explicitly signaled by indicating, in addition to the country of citizenship language (i.e. German, Dutch or Spanish, as applicable), a second mother tongue, e.g. 'Russian (mother tongue)'.

Based on the country of ancestry of the applicant, we differentiate four large regions of ancestry: Europe and the USA,[15] Latin America and the Caribbean,[16] Middle East and North Africa (MENA)[17] and Asia.[18] These four regions include applicants of the four phenotypic groups (White, DSC, ASIN and Black). In addition to these four regions, we include applicants (whose parents come) from sub-Saharan Africa.[19] Table 1 displays the distribution of phenotypic groups across regions of ancestry (for the distribution of phenotypic groups across specific countries of ancestry, see Supplementary Table A3).

Callback

The dependent variable is a binary variable differentiating positive (signal of interest) from negative (no signal of interest) employer response. Employers can signal interest in three ways: First, the employer invites the job candidate for an interview. Second, the job applicant receives a 'pre-invitation': the candidate has passed an early selection process, but s/he is not (yet) formally invited for an interview/meeting. Third, employers sometimes request additional information or ask to be called back. The category 'No signal of interest' can have the following reasons: employers explicitly reject the job applicant, there may be no response at all 12 weeks after application, or the only response by the employer is a mere confirmation of receipt. Table 2 displays callback rates by phenotypic groups for different subsamples. Note that these descriptive statistics do not account for variation in applicants' ancestry.

Controls

We control for differences in the occupational distribution by including occupation fixed-effects. We also include applicants' gender because our pre-tests showed female photos are

14 Name recognition was tested via online plausibility surveys carried out in all three countries of this study. These tests showed perceivers are typically better at matching names to regions of ancestry than to specific countries (an overview of the name selection procedure can be found in Lancee *et al.*, 2019b).

15 The Europe and US region includes applicants whose parents come from Europe (i.e. Albania, Belgium, Bosnia, Bulgaria, Denmark, France, Germany, Greece, Italy, Macedonia, Netherlands, Poland, Portugal, Norway, Romania, Russia, Spain, Sweden, Ukraine and UK) and the USA.

16 The Latin America and the Caribbean region includes Mexico, Dominican Republic, Trinidad and Tobago, Suriname, Antilles and Ecuador (the latter overrepresented in Spain).

17 The MENA region includes Iran, Iraq, Lebanon (overrepresented in Germany), Morocco (overrepresented in Spain and the Netherlands) and Turkey (overrepresented in Germany and the Netherlands).

18 The Asian region includes China, India, Indonesia, Japan, Malaysia, Pakistan, Philippines and Vietnam.

19 The sub-Saharan African region includes Ethiopia, Uganda, Nigeria and South Africa.

Table 1 Distribution of phenotypic groups across regions of ancestry

Ethnic ancestry (regions)	Phenotypic groups				No photo	Total
	'WHITE' (White 1, White 2, White 3)	'DSC' (DSC)	'ASIN' (ASIN 1, ASIN 2)	'BLACK' (Black 1, Black 2)		
						
Europe and USA	3625	1021	87	164	1507	6404
	56.61	15.94	1.36	2.56	23.53	100.0
	72.14	49.71	5.95	11.97	50.44	49.65
MENA	996	716	401	410	841	3364
	29.61	21.28	11.92	12.19	25.00	100.0
	19.82	34.86	27.43	29.93	28.15	26.08
Latin America and Caribbean	222	178	187	259	183	1029
	21.58	17.30	18.17	25.17	17.78	100.0
	4.42	8.67	12.79	18.91	6.12	7.98
Asia	182	139	787	41	326	1475
	12.34	9.42	53.35	2.78	22.10	100.0
	3.62	6.77	53.83	2.99	10.91	11.43
Africa	0	0	0	496	131	627
	0.0	0.0	0.0	79.11	20.89	100.00
	0.0	0.0	0.0	36.20	4.38	4.86
Total	5025	2054	1462	1370	2988	12 899
	38.95	15.92	0.0	0.0	23.16	100.0
	100.0	100.0	100.0	100.0	100.0	100.0

Notes: European and US ancestry includes France, Germany, Greece, Italy, the Netherlands, Norway, Spain, the UK, the USA, Belgium, Portugal, Albania, Bulgaria, Poland, Romania, Russia, Bosnia, Macedonia and Ukraine; MENA ancestry includes Egypt, Iran, Iraq, Lebanon, Morocco and Turkey; Latin America and the Caribbean ancestry includes Mexico, Dominican Republic, Trinidad and Tobago, Suriname, Antilles and Ecuador; Asian ancestry includes China, India, Indonesia, Japan, Pakistan, South Korea, Vietnam, Malaysia and Philippines; and African ancestry includes Ethiopia, Nigeria, Uganda and South Africa.

Table 2 Descriptives: mean raw callback rates by phenotypic groups

	All (1)	White (2)	DSC (3)	ASIN (4)	Black (5)	No photo (6)	White/Black ratio (7)	White/Black odds ratio (8)	White–Black percent difference (*P*-value) (9)
All	36.43	39.83	33.66	32.10	28.30	38.51	1.41	1.68	11.53
	[12.783]	[4.979]	[2.038]	[1.455]	[1.135]	[2.947]			[0.0000]
Male	35.04	37.39	34.32	29.99	26.05	37.98	1.44	1.70	11.34
	[6.178]	[2.495]	[947]	[697]	[641]	[1.398]			[0.0000]
Female	37.73	42.27	33.09	34.04	30.29	38.99	1.40	1.69	11.98
	[6.605]	[2.484]	[1.091]	[758]	[723]	[1.549]			[0.0000]
Germany	49.57	54.40	46.88	42.61	47.38	46.39	1.15	1.32	7.02
	[3.161]	[1.399]	[544]	[467]	[363]	[388]			[0.017]
The Netherlands	46.23	54.91	47.41	43.41	40.39	41.89	1.36	1.80	14.52
	[4.402]	[1.253]	[483]	[334]	[255]	[2.077]			[0.0000]
Spain	20.21	22.95	19.98	18.81	14.88	17.63	1.54	1.70	8.07
	[5.220]	[2.327]	[1.011]	[654]	[746]	[482]			[0.0000]

Notes: The table reports, for the entire sample and different subsamples of sent applications, the callback rates for: All applicants (Column 1), White applicants (Column 2), DSC applicants (Column 3), ASIN applicants (Column 4), Black applicants (Column 5) and applicants with no photograph (Column 6). It also reports the White/Black callback ratio (Column 7) and the White/Black odds ratio (Column 8). In brackets in each cell is the number of applications sent in that cell. Finally, Column 9 reports the *P*-values for a test of proportion testing the null hypothesis that the callback rates are equal across White and Black applicants.

generally rated more positively than male ones (see Supplementary Table A1). The job candidate was randomly assigned a male or female first name and the gender was mentioned explicitly in the CV.

3.2 Analytical strategy

We investigate appearance-based racial discrimination in Europe by fitting three main statistical models. In the first model, we estimate callback probabilities for different combinations of phenotype and region of ancestry. This full model, which includes Black applicants of Sub-Saharan African descendants as an additional benchmark treatment, allows us to estimate the impact of phenotypic variation within each region of ancestry.

Our second statistical model estimates the difference in callback rates for each non-White phenotype as compared to the White phenotype across the four different regions of ancestry. Sub-Saharan African descendants are dropped from this model because we do not have full phenotypic variation for this region of ancestry. The second model is therefore a (non-nested) reduced model, which yields an estimate of appearance-based racial discrimination *net of ancestry* discrimination in each of the analyzed countries. The reduced model thus assumes that applicants' phenotype triggers a similar response in employers regardless of applicants' region of ancestry and hence, we argue, serves as a test for the hypothesis that racial appearance has an *independent* effect on employers' hiring decisions.[20]

The last model takes the reduced model as the basis and investigates country differences in appearance-based racial discrimination. We estimate an interaction term between phenotype and country of the experiment, while keeping applicants' region of ancestry constant. This allows us to test whether German, Dutch and Spanish employers are equally responsive to variation in applicants' phenotype, and hence to test for our expectation that Spanish employers could be more likely to discriminate against non-White minorities—an expectation that turns out to be wrong.

3.3 Comparing discrimination estimates across countries

To model employers' callback probabilities under different treatments, we estimate logistic regression models and report odds ratios (ORs). ORs provide a measure that is independent of the marginal distribution of callbacks and thus particularly well-suited for comparing employers' responses across countries with different average callback rates.[21] Illustrative predicted marginal callback probabilities for the main models are discussed in the text and shown in full in the Supplementary Appendix. To evaluate whether the reduced model provides a better statistical representation of the data structure than the full model, we draw on the Akaike Information Criterion (AIC) statistic, which is especially fitting for comparing between non-nested models that differ in the number of parameters (see, e.g. Bailey *et al.*, 2014).[22] In

20 Both the full model and the reduced model include a non-photograph condition. We note not attaching a photograph to the CV is heavily penalized in all three countries and with similar intensity.

21 Average callback rates differ substantially across the countries of this study. Whereas the Netherlands and Germany show average callback rates of 46% and 50%, respectively, the average callback rate in Spain is only 20%.

22 The AIC penalizes additional parameters less heavily than the Bayesian Information Statistic (BIC), which is also widely used for non-nested model selection. This means that the AIC provides the

line with the specialized literature, we consider only absolute differences above 10 points between the reduced model and the full model as providing sufficiently strong statistical evidence in favor of the former (see Burnham and Anderson, 2004, p. 271; Bailey *et al*., 2014). Last, to provide a statistical test of country differences in net appearance-based racial discrimination, we use a linear probability model (LPM). We use this specification to overcome the well-known limitations of testing for interactions in non-linear probability models (Ai and Norton, 2003). We note, however, all the findings reported in this study are fully robust to using LPMs throughout.

4. Findings

4.1 The combined effect of phenotype and ancestry: full models

Table 3 shows the combined effect of phenotype and region of ancestry in triggering employers' responses to our fictitious applicants in Germany, the Netherlands and Spain, as well as in a pooled model with country fixed effects. Results are presented graphically in Figure 2 (pooled model) and Figure 3 (country models). Before commenting on the results of these full models, it is important to bear in mind that combining racial appearance and ancestry takes a toll on statistical power thus increasing the probability of type-II error, particularly in small-n cells within country models. Results should therefore be interpreted cautiously.

Both the pooled model as well as the models for Germany and the Netherlands suggest that phenotype plays a similar role in eliciting employers' responses across applicants of different regions of ancestry, which would constitute preliminary evidence of *independent effects* of phenotype on employers' hiring decisions. At first sight, the country that most closely resembles this 'pure' phenotypic discrimination scenario seems to be the Netherlands, where we find what appears to be a color hierarchy (White–DSC–ASIN–Black) in employers' responses to applicants of all ethnic ancestries with the sole exception of applicants of European–US ancestry. Only in this latter group of applicants we fail to find evidence that the Black phenotype is penalized by employers, thus breaking the common pattern found for all other ancestry regions. Note, however, that given the small sample size of this group ($n = 29$), we cannot rule out the possibility that we are simply failing to capture a true effect due to lack of statistical power. Germany shows a very similar pattern to that found in the Netherlands. The clearest deviation from a common color hierarchy pattern in this case is Black German applicants of Latin American/Caribbean ancestry ($n = 70$), the only Black group in Germany for which the null hypothesis of no discrimination cannot be ruled out. Overall, the data for Germany and the Netherlands strongly suggest that applicants' phenotype can be an important source of discrimination, which seems to operate with a logic of its own. This interpretation is further confirmed by comparing information criterion statistics between the full model and the reduced model. The AIC statistic for the reduced model is 16 points lower in Germany and 48 points lower in the Netherlands than for the full model (see Supplementary Table A6), which we interpret as strong evidence that in these two societies racial appearance acts indeed as an independent trigger of discriminatory behavior, which operates *in addition* to ancestry discrimination.

most conservative (stringent) test for the reduced model (see Burnham and Anderson, 2004, p. 271). Both BIC and AIC statistics are shown in Supplementary Table A6.

Table 3 Full models: the combined effect of phenotype and ancestry

Employer's interest	Pooled Odds ratio	Germany Odds ratio	The Netherlands Odds ratio	Spain Odds ratio
European and US ancestry				
DSC	0.835**	0.708**	0.779*	0.929
	[0.0649]	[0.0983]	[0.110]	[0.141]
ASIN	0.742	0.522	0.599	1.445
	[0.175]	[0.211]	[0.223]	[0.590]
Black	0.636**	0.712	0.961	0.441**
	[0.115]	[0.200]	[0.375]	[0.153]
Middle East and North African ancestry				
White	0.802***	0.796*	0.737**	0.804
	[0.0659]	[0.102]	[0.111]	[0.130]
DSC	0.614***	0.723**	0.561***	0.530***
	[0.0603]	[0.117]	[0.101]	[0.0961]
ASIN	0.590***	0.571***	0.506***	0.748
	[0.0709]	[0.0950]	[0.116]	[0.198]
Black	0.528***	0.416**	0.463***	0.572***
	[0.0713]	[0.174]	[0.106]	[0.107]
Latin American and Caribbean ancestry				
White	0.892	1.108	0.887	0.857
	[0.142]	[0.421]	[0.382]	[0.174]
DSC	0.671**	0.464*	0.924	0.728
	[0.127]	[0.215]	[0.530]	[0.165]
ASIN	0.524***	0.376	0.620	0.534***
	[0.103]	[0.233]	[0.286]	[0.128]
Black	0.746**	0.913	0.496*	0.754
	[0.110]	[0.228]	[0.192]	[0.162]
Asian ancestry				
White	0.922	1.076	0.798	0.856
	[0.152]	[0.298]	[0.233]	[0.260]
DSC	0.937	0.744	0.734	1.330
	[0.177]	[0.222]	[0.271]	[0.404]
ASIN	0.685***	0.642***	0.613***	0.770*
	[0.0606]	[0.0947]	[0.102]	[0.116]
Black	0.567	0.295	0.318	1.064
	[0.225]	[0.239]	[0.261]	[0.561]
Other				
Sub-Saharan African Black	0.614***	0.746*	0.566**	0.436***
	[0.0675]	[0.118]	[0.134]	[0.101]
No photo condition	0.608***	0.699***	0.539***	0.638***
	[0.0352]	[0.0866]	[0.0446]	[0.0939]
Observations	12 783	3161	4402	5220

Notes: Logistic callback probability estimates for country-specific and pooled models, reference is White applicants of European/US ancestry.

Robust standard errors in brackets.

Models include additional controls for gender of the applicant, type of occupation, country of experiment and religion signal (Muslim). Intercept not shown. Catalan names in Catalonia are weighted in the Spanish model to account for their majority status in this region.

***$P < 0.01$; **$P < 0.05$; *$P < 0.10$.

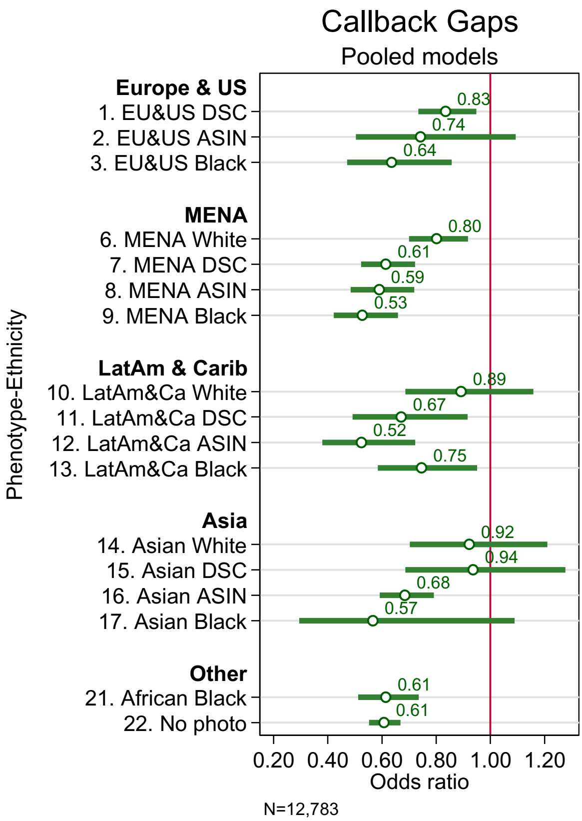


Figure 2 Phenotype–ancestry discrimination estimates (OR), pooled model, reference is White applicants of European–US ancestry.
Source: Table 3.

The additive effect of ethnic-ancestry discrimination and appearance-based discrimination is sizable in Germany and the Netherlands. According to our estimates, Black applicants of MENA ancestry in Germany and the Netherlands have between *(*(1−(1/OR))×100=) 117% (the Netherlands) to 138% (Germany) lower odds (between 32% and 38% lower marginal probabilities) to receive a callback than identical White applicants of European–US ancestry, while White applicants of MENA ancestry have 'only' between 25% (Germany) and 35% (the Netherlands) lower odds (i.e. between 10% and 13% lower marginal probabilities, respectively) to receive a call-back than identical White applicants of European–US ancestry. Looking within the MENA ancestry group, which as explained above is a particularly relevant group in the European context, we find substantial differences in callback probabilities by racial appearance. According to our estimates, the callback rate for White MENA applicants is around 50% in both countries, while the callback rate for identical Black applicants goes down to roughly 36%. This amounts to a decline of around 25% in marginal callback probabilities and means that if a White MENA applicant has to send roughly two applications to get a callback in either the Netherlands or

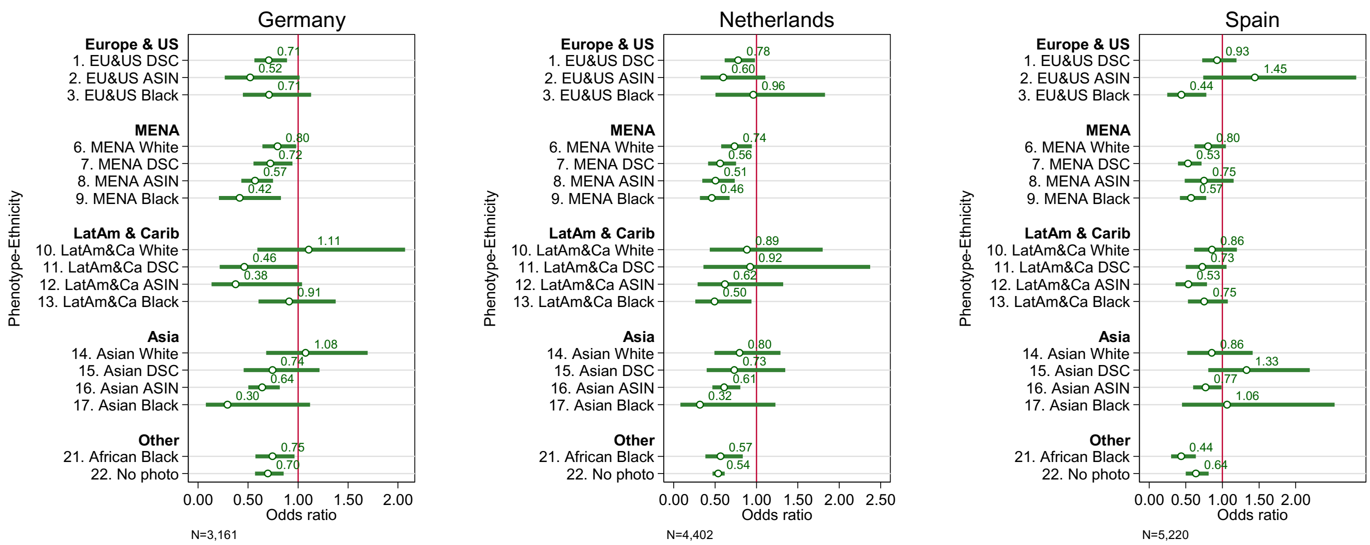


Figure 3 Phenotype–ancestry discrimination estimates (OR) in Germany, the Netherlands and Spain, male and female applicants pooled, reference is White applicants of European–US ancestry.

Source: Table 3.

Germany, an identical Black MENA applicant would have to send three (see Supplementary Table A4 for other marginal probabilities). These findings show appearance-based racial discrimination operates *within* the MENA ancestry group leading to substantial differences in employment opportunities. Applicants' phenotype therefore seems to act as an additional barrier to employment, triggering discriminatory behavior by itself.

Phenotype seems to have a somewhat different and more complex effect in Spain. Not only discrimination estimates seem at first sight lower across regions of ancestry in this country, contradicting our original expectations, but it also seems more difficult to find consistent color hierarchies (see Figure 3). Instead, the response pattern found among the over 5000 firms tested in Spain suggests employers' discriminatory behavior could be triggered by particular combinations of phenotype and ancestry; specially, combinations where phenotype is more prototypical of the region of ancestry. These prototypical combinations include MENA ancestry applicants with DSC and Black phenotypes, Latin American ancestry applicants with ASIN/Indigenous phenotype, Asian ancestry applicants with ASIN phenotype ($P = 0.08$) and African ancestry applicants with Black phenotype. One plausible interpretation of these findings is that phenotype in Spain, rather than having an independent effect on employers' callback probabilities, as it seems to be the case in both Germany and the Netherlands, acts as a reinforcing signal of ancestry, increasing its salience in the cases where phenotypes are most prototypical. Consistent with this interpretation, we note the reduced model, which assumes independent effects of phenotype, failed to pass the AIC test in the Spanish experiment ($AIC_{full} - AIC_{reduced} = 5$) (see Supplementary Table A6).

The combined effect of phenotype and ancestry can also lead to substantial discrimination in Spain. According to our estimates, DSC and Black applicants of MENA ancestry as well as Black applicants of Sub-Saharan ancestry show 89%, 75% and 127% lower callback odds respectively (38%, 34% and 47% lower marginal predicted probabilities) when compared to White applicants of European–US ancestry. While White applicants of European–US ancestry have a callback probability of 23% in Spain, DSC and Black applicants of MENA ancestry and Black African-ancestry applicants have an averaged callback probability of around 14% (see other predicted margins of phenotype in Supplementary Table A4 and Supplementary Figure A1). In contrast, the callback rate for White applicants of MENA ancestry is 20%. This means there are also significant differences in callback probabilities by phenotype *within* the MENA ancestry group in Spain. Specifically, Spanish MENA-ancestry applicants with a DSC phenotype have 29% lower marginal probabilities to receive a callback than identical MENA-ancestry applicants with a White phenotype. These estimates, together with those found for Asian applicants and Latin American applicants with ASIN/Indigenous phenotypes, are in line with a prototypical phenotype interpretation of the Spanish data. Yet, it must be noted, one exception from this prototypicality pattern is Black applicants of European–US ancestry ($n = 81$), who show 47% lower marginal callback probabilities than White applicants of the same region of ancestry in Spain.

4.2 Net racial discrimination estimates: reduced models

Table 4 shows the results of fitting the reduced models to each of the three countries of the experiment, as well as to the three countries pooled. As explained above, reduced models aim to capture *net* appearance-based racial discrimination effects. ORs for each racial group

Table 4 Reduced models: net phenotypic discrimination

Employer's interest	Pooled Odds ratio	Germany Odds ratio	The Netherlands Odds ratio	Spain Odds ratio
Applicant's phenotype (Ref.-> White)				
DSC	0.826***	0.765***	0.795**	0.908
	[0.0491]	[0.0789]	[0.0885]	[0.0882]
ASIN	0.732***	0.677***	0.697**	0.834
	[0.0577]	[0.0874]	[0.0995]	[0.120]
Black	0.723***	0.707*	0.680**	0.771*
	[0.0664]	[0.132]	[0.121]	[0.107]
Region of ancestry (Ref.-> Europe and USA)				
Middle East and North Africa	0.730***	0.811**	0.641***	0.755**
	[0.0410]	[0.0797]	[0.0562]	[0.0861]
Latin America and the Caribbean	0.850*	1.055	0.782	0.818*
	[0.0710]	[0.202]	[0.130]	[0.0977]
Asia	0.916	0.914	0.861	0.964
	[0.0689]	[0.124]	[0.0988]	[0.142]
Country of field experiment (Ref.-> Spain)				
Germany	3.783***			
	[0.203]			
The Netherlands	3.497***			
	[0.184]			
Constant	0.511***	1.791***	2.080***	0.458***
	[0.0386]	[0.247]	[0.251]	[0.0573]
Observations	12 161	2940	4226	4995
Pseudo-R^2	0.10	0.02	0.05	0.04

Notes: Logistic callback probability estimates (OR) for pooled and country-specific models.
Robust standard errors in brackets.
Models include additional controls for gender of the applicant, type of occupation and religion signal (Muslim).
$^{***}P<0.01$; $^{**}P<0.05$; $^{*}P<0.1$.

are presented graphically in Figure 4 (for predicted margins see Supplementary Table A5 and Supplementary Figure A2).

The first model in Table 4 is the pooled model, which estimates the cross-ancestry cross-country odds of receiving a callback by employer for applicants of three different minority phenotypes, DSC, ASIN and Black, when compared to White applicants. The pooled model shows applicants' phenotype adds to region of ancestry as independent sources of discrimination in Europe. According to the estimates of the pooled model, applicants with a DSC phenotype would have ((1−(1/0.83))×100=) 20% lower odds (10% lower marginal probabilities) to get a callback than identical White applicants, while applicants with an ASIN phenotype and applicants with a Black phenotype would have roughly 40% lower odds (16% lower marginal probabilities) to receive a callback than identical Whites. Expressed in callback probabilities, we note our callback rate estimate for 'typical' White applicants (across the three countries studied) is 41%, while the estimated callback rate for identical Black applicants is roughly 34% (see predicted margins on Supplementary Table A5). This

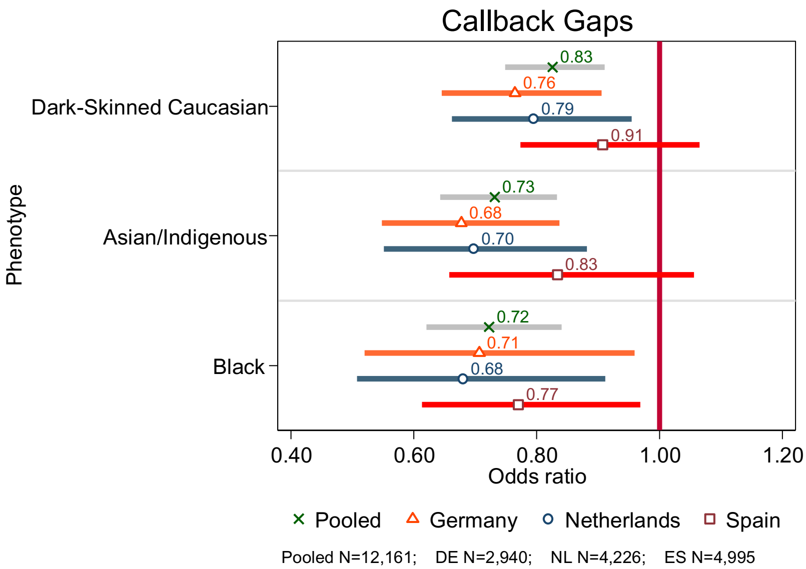


Figure 4 Cross-ancestry phenotypic discrimination estimates. Applicants of parents born in Europe–USA, MENA, Latin American–Caribbean and Asian Countries (region of ancestry fixed effects), male and female applicants pooled, reference is White phenotype.
Source: Table 4.

model provides the best explanation of the pooled data, as shown by the AIC statistic ($AIC_{full} - AIC_{reduced} = 45$) (see Supplementary Table A6).

4.3 Testing country differences

The ORs presented in the pooled model are cross-country estimates and hence assume net appearance-based racial discrimination has similar intensity across the three countries of our study. In line with our previous discussion of the full models, country-specific reduced models yield a lower estimate for net appearance-based racial discrimination in Spain, when compared to either Germany or the Netherlands. In fact, in Spain, we fail to find evidence of net discrimination against applicants with the DSC or the ASIN phenotypes, while the estimate for net discrimination against the Black phenotype is comparatively low and only significant at the 90% level. According to our estimates, the net callback rate for White applicants in Spain is around 22%, while the callback rate for identical Black applicants goes down to 18%. This means Black applicants in Spain have *(*(1−(1/0.77))×100=) 30% lower odds to receive a callback from employers than identical White applicants net of region of ancestry, which seems a lower penalty than that found in the German and the Dutch experiments. To test this country difference formally, Table 5 below shows the results of fitting a pooled reduced model where country of the experiment is interacted with applicants' phenotype, while keeping region of ancestry constant. This model shows the difference in net racial discrimination between Spain, on the one hand, and the Netherlands and Germany, on the other, is indeed statistically significant and robust to different

Table 5 Lower net phenotypic discrimination in Spain confirmed with three different measures, interacted linear probability models

	M1 LPM	M2 LPM	M3 LPM
Employer interest			
Racial groups (*Ref. White*)			
DSC	−0.0655***		
	[0.0166]		
ASIN	−0.0989***		
	[0.0205]		
Black	−0.101***		
	[0.0270]		
Region of ancestry (*Ref. Europe and USA*)			
Middle East and North Africa	−0.0774***	−0.0789***	−0.0810***
	[0.0119]	[0.0118]	[0.0117]
Latin America and the Caribbean	−0.0251	−0.0230	−0.0321*
	[0.0182]	[0.0181]	[0.0177]
Asia	−0.0188	−0.0237	−0.0282*
	[0.0184]	[0.0162]	[0.0160]
Country of field experiment (*Ref. Germany and the Netherlands*)			
Spain	−0.303***	−0.329***	−0.304***
	[0.0130]	[0.0192]	[0.0130]
Spain × DSC	0.0515**		
	[0.0237]		
Spain × ASIN	0.0733***		
	[0.0270]		
Spain × Black	0.0659*		
	[0.0340]		
Racial gradient (White–DSC–ASIN–Black)		−0.0420***	
		[0.00722]	
Spain × Racial gradient		0.0303***	
		[0.00925]	
Non-White			−0.0806***
			[0.0137]
Spain × Non-White			0.0612***
			[0.0187]
Observations	9343	9343	9343
R^2	0.125	0.125	0.125

Robust standard errors in brackets.
Models include additional controls for gender of the applicant and type of occupation. Intercept not shown.
$^{***}P < 0.01$; $^{**}P < 0.05$; $^{*}P < 0.1$.

operationalizations of the phenotype variable. While, according to the classic intergroup contact and competitive threat theories discussed above, Spain should provide the most fertile ground for the proliferation of discriminatory behavior, Spain is actually the country with the lowest levels of net appearance-based racial discrimination of the three countries studied—as well as the country with the lowest levels of net ancestry discrimination (see Supplementary Table A7). We discuss this unexpected finding in the concluding section.

At first sight, the Netherlands appears as the country with the highest estimates of net phenotypic discrimination, which is the largest for Black applicants. The estimated net callback rate for White applicants in the Netherlands is 54%, while it is 45% for identical Black applicants. Black Dutch applicants have $((1-(1/0.68))\times100=)$ 47% lower odds to receive a callback from employers than identical White applicants. In Germany, Black applicants have $((1-(1/0.71))\times100=)$ 41% lower odds to receive a callback than identical White applicants (see Supplementary Table A5 for predicted margins). Differences in net phenotypic discrimination between the Netherlands and Germany are, however, not statistically significant. We also note there are no statistical differences in appearance-based racial discrimination by gender in none of the three countries studied (results not shown).

4.4 Robustness tests

To check the robustness of our findings, we have carried out a host of different tests. First, we have checked whether our discrimination estimates depended on the different distribution of countries of ancestry across the three countries studied. As explained above, our standardized research design includes applicants from 31 countries of ancestry that are common to all three countries of the experiment plus seven ancestries of special interest in each country that are not (see Supplementary Table A2). One concern with this design is that the observed country differences in discrimination estimates could be artificially driven by differences in the distribution of specific ancestries (even though our plausibility tests clearly showed native participants are better at matching applicants' names to large regions rather than to specific countries of ancestry and our results are robust to clustering standard errors by country of ancestry). To address this concern, we have re-estimated all models using only applicants for the 31 common ancestries and results clearly hold.[23] We have also checked for the robustness of our full models to (a) excluding children of Eastern European parents from the European ancestry category, (b) using White children of native parents alone as the reference category and (c) removing all children of native parents from the analytical sample. The main patterns within and across countries hold regardless of the chosen reference category (even though statistical power logically diminishes with the reduction in sample size). Finally, we have checked whether our results are robust to removing female applicants with a headscarf from the analytical sample, this being a specific treatment for Muslim women that has been explored elsewhere (see Fernández-Reino *et al.*, 2022). Results again hold. We stress the difference in the intensity of net appearance-based racial discrimination found between Spain, on the one hand, and Germany and the Netherlands, on the other, is robust to all these tests. All our findings are also robust to using LPMs instead of logistic regression to model employers' callbacks (all tests are available upon request).

23 Note this test also deals with the potential biasing impact of having a different distribution of photographs within each large phenotypic group within each region of ancestry, which happens when using the full set of ancestries. An alternative way of testing for this potential bias is restricting the analysis to only the four phenotypes that are present in all four regions (i.e. White 3, DSC, ASIN 2 and Black 1). Results are also robust to this alternative test (available upon request).

5. Discussion and conclusions

This study presents the results of the first large-scale comparative field-experiment on appearance-based racial discrimination conducted in Europe. We sent almost 13 000 fictitious job applications to an equal number of real vacancies advertised online in three European Union countries, Germany, the Netherlands and Spain. We exploited randomized variation in racial appearance across four regions of ancestry and this allowed us to investigate the combined and net effects of applicants' ancestry and applicants' phenotype in real job-selection processes. We found strong evidence that applicants' racial appearance triggers discriminatory behavior in all three countries studied. To put it bluntly, many immigrant descendants in Europe are discriminated against because they have visibly atypical (i.e. non-White) phenotypes. As the pool of second-generation applicants entering employment increases, the number of new Europeans at risk of suffering appearance-based racial discrimination is on the rise. Considering these findings, the neglect of 'race' that has long characterized scholarship in mainland Europe would seem hard to justify.

But are European employers using racial categories in and of themselves when discriminating minority phenotypes or do minority phenotypes act as blurred signals of 'otherness' that reinforce ethnic ancestry signals? In other words, are European employers purely *color-racist* or are they *ethno-racist*? Although the difference might seem subtle, we believe answering this question has great substantive import. To our knowledge, this question has never been addressed before in correspondence-test studies, among other things, because—with a handful of exceptions—previous research has typically confounded ethnic background and phenotype in one single treatment, i.e. applicant's name. Our unique research design allowed us to address, or at least begin to address, this crucial question by investigating the interplay between applicants' phenotype and their ethnic ancestry in eliciting employers' responses.

We found that Black applicants tend to receive the lowest callback rates on average in the country-pooled models net of applicants' region of ancestry (while White applicants receive the highest), although it must be noted callback rates for Black applicants are only marginally different from those found for other non-White groups (particularly ASIN). To our knowledge, the net discrimination estimates reported in this study are the first cross-country comparable *net* appearance-based racial discrimination estimates ever reported in a correspondence test. These net estimates are lower than the name-based racial discrimination estimates typically found in the US literature for African Americans (see, e.g. Bertrand and Mullainathan, 2004; Pager and Shepherd, 2008, Quillian *et al.*, 2017). Interestingly, our net estimates for visible phenotypes are similar in size to the estimates for the MENA ancestry treatment in the reduced models (see Table 4). This strongly suggests that phenotype is indeed a very important omitted variable in European scholarship. By overlooking phenotype, previous scholarship might have likely overstated the role of culture and religion as drivers of discrimination in Europe.

The reduced model provides the best fit to both the German and the Dutch data. This suggests applicants' phenotype operates as an *independent* trigger of hiring discrimination in these countries, where phenotypic discrimination seems to add to ancestry discrimination in hindering access to employment. Phenotypic discrimination estimates found for German and Dutch applicants of non-European ancestry are severe and in line with—if not larger than—the (name-based) estimates typically reported for African American applicants in the USA (see Quillian *et al.*, 2017).[24]

No such clear independent effects of phenotype are found in the Spanish data, however, where the AIC provides little support for the reduced model (when compared to the full model). The lack of (evidence of) *independent* effects of phenotype should, however, not be mistaken for lack of effects. What the results of the Spanish experiment suggest, instead, is a less direct effect of applicants' appearance on employers' responses—i.e. a more complex interplay of phenotype and ancestry. Rather than reacting to applicants' phenotype alone, Spanish employers seem to react to particular combinations of phenotype and ancestry, especially, we have tentatively argued, to those where phenotype is prototypical of the region of ancestry. In a manner akin to a (hypothetical) typical costume, prototypical phenotypes could act as reinforcing signals of ancestry, thus increasing ethnic salience in applicants' résumés. While this tentative interpretation is in line with laboratory research on phenotypic prototypicality (PP) (Maddox, 2004), it must be noted we also found significantly lower callback probabilities for Black applicants of European–US ancestry in Spain. This latter finding seems admittedly harder to reconcile with a simple PP effect, which leaves the question of mechanisms open for future research.

Given that we fail to find a fully independent effect of phenotype on Spanish employers' callbacks, it is not surprising that our net phenotypic discrimination estimates for this country are significantly lower than those found in Germany or the Netherlands. Finding the lowest levels of net phenotypic discrimination in Spain may, however, be considered surprising from a theoretical standpoint for, as discussed above, Spain experienced a very rapid increase in the size of its migrant population, which was subsequently followed by an equally rapid and very intense deterioration of the economic environment. We have argued this combination should provide a particularly fertile ground for the proliferation of discriminatory behavior against new visible minorities. Contrary to this expectation, we found Spanish employers are less likely to engage in direct phenotypic discrimination than their German or Dutch counterparts.[25]

While at this stage we can only speculate about the reasons for this apparently puzzling finding, we suspect there might be interesting differences in the colonial legacies of the three countries studied that are worth exploring. For example, the co-existence of black slavery and indentured servitude (of British Indians and Javanese) in the Dutch Antilles, together with the institution of two-class citizenship, might have helped generate brighter racial hierarchies in the Netherlands—particularly when compared to the more fluid ethno-racial boundaries imposed by Spanish colonial powers in the Americas (see Cope, 1994; Mörner, 1967; Weiner, 2014; Wekker, 2016; see also Telles and the Project on Ethnicity and Race in Latin America (PERLA), 2014). We also note that color-based racism was not a cornerstone concept for Spanish Fascism, whereas it was indeed quintessential to German (and Dutch) Nazism. Finally, we note there are also significant differences in the degree of salience of

24 According to Quillian *et al.*'s (2017) meta-analysis of all field-experiments on racial discrimination carried out in the USA since 1989 (n=24), the average estimate for the White-to-African American callback ratio (CBR) is 1.45 (i.e. African American have to send 45% more applications than identical Whites to receive a callback by employers).

25 Additional analyses suggest Spanish employers are also less likely to engage in direct ethnic discrimination (available upon request).

immigration in contemporary politics between Spain, on the one hand, and the Netherlands and Germany, on the other. In contrast to the latter two countries, Spain did not have an openly anti-immigrant party in parliament up until the April general election of 2019 (i.e. after our fieldwork was completed). To the extent that historical legacies and contemporary political dynamics play a crucial role in ethno-racial boundary-making processes (Wimmer, 2013; Polavieja and Fischer-Souan, 2022), these country differences should deserve special attention in future research.

This study fills an important gap in the stratification and immigrant incorporation literature in Europe. We hope it also inaugurates a new avenue of empirical research on 'race'. One that not only sheds light on the mechanisms of racial stratification in increasingly diverse European societies, but also on the complex interplay of phenotype and ethnicity *everywhere*.

Acknowledgements

The leading author presented previous versions of this article at: the Interdisciplinary Experimental Working Group, European University Institute, Florence; the Nuffield College Sociology Seminar, University of Oxford; the Instituto de Empresa, Madrid; the Institute of Public Goods and Policies, Spanish National Research Council, Madrid; the Deaton Review Workshop on 'Immigration, Race and Inequality', Institute of Fiscal Studies & CReAM, University College London; the International Inequalities Institute, London School of Economics; the DIGCLASS Seminar Series, European Commission and the MaD Colloquium at the WZB Berlin Social Science Center. He wishes to thank all participants in these venues for their valuable input and, specially, Richard Breen, Christian Dustmann, Klarita Gërxhani, Stephen Jenkins, Irene Pañeda, Lucinda Platt and Leire Salazar.

Supplementary material

Supplementary material is available at *Socio-Economic Review Journal online*.

Funding

This research has been funded by the Horizon 2020 Program of the European Commission as part of the project 'Growth, Equal Opportunities, Migrations and Markets' [grant number H2020 649255], as well as by the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033/) as part of the project 'Pushing the Boundaries of Research on Ethno-Racial Discrimination in Hiring' [grant number PID2020-119558GB-I00].

References

Adida, C. L., Laitin, D. and Valfort, M. A. (2016) *Why Muslim Integration Fails in Christian-Heritage Societies*, Cambridge, MA, Harvard University Press.

Ahmed, A. M., Andersson, L. and Hammarstedt, M. (2013) 'Are Gay Men and Lesbians Discriminated against in the Hiring Process?', *Southern Economic Journal*, **79**, 565–585.

Ai, A. and Norton, E. C. (2003) 'Interaction Terms in Logit and Probit Models', *Economics Letters*, **80**, 123–129.

Alba, R. (2005) 'Bright Vs. Blurred Boundaries: Second-Generation Assimilation and Exclusion in France, Germany, and the United States', *Ethnic and Racial Studies*, **28**, 20–49.

Allport, G. W. (1954) *The Nature of Prejudice*, Cambridge, MA, Perseus Books.

Bailey, S., Saperstein, A. and Penner, A. M. (2014) 'Race, Color and Income Inequality across the Americas', *Demographic Research*, **31**, 735–756.

Balibar, É. (1991) 'Is There a Neo-Racism?'. In Balibar, E. and Wallerstein, I. (eds) *Race, Nation, Class: Ambiguous Identities*, London, Verso. pp. 17–28.

Banton, M. (2012) 'John Rex's Main Mistake', *Ethnic and Racial Studies*, **3**, 1369–1381.

Bertrand, M. and Mullainathan, S. (2004) 'Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination', *American Economic Review*, **94**, 991–1013.

Bertrand, M. and Duflo, E. (2017) 'Field Experiments on Discrimination'. In Banerjee, A. V. and Duflo, E. (eds) *Handbook of Economic Field Experiments*, Vol. **1**, Amsterdam, North-Holland. pp. 309–393.

Blalock, H. (1967) *Toward a Theory of Minority Group Relations*, New York, NY, Wiley.

Bonilla-Silva, E. (1997) 'Rethinking Racism: Towards a Structural Interpretation', *American Sociological Review*, **62**, 465–480.

Bourdieu, P. and Wacquant, L. (1999) 'On the Cunning of Imperialist Reason', *Theory, Culture & Society*, **16**, 41–58.

Brubaker, R. (2009) 'Ethnicity, Race and Nationalism', *Annual Review of Sociology*, **35**, 21–42.

Burnham, K. and Anderson, D. (2004) 'Multimodel Inference: Understanding AIC and BIC in Model Selection', *Sociological Methods & Research*, **33**, 261–304.

Charles, M. and Guryan, J. (2011) 'Studying Discrimination: Fundamental Challenges and Recent Progress', *Annual Review of Economics*, **3**, 479–511.

Connor, P. and Koenig, M. (2015) 'Explaining the Muslim Employment Gap in Western Europe: Individual-Level Effects and Ethno-Religious Penalties', *Social Science Research*, **49**, 191–201.

Cope, D. (1994) *The Limits of Racial Domination: Plebeian Society in Colonial Mexico City, 1660–1720*, Madison, University of Wisconsin Press.

Derous, E., Pepermans, R. and Ryan, A. M. (2017) 'Ethnic Discrimination during Résumé Screening: Interactive Effects of Applicants' Ethnic Salience with Job Context', *Human Relations*, **70**, 860–882.

Di Stasio, V., Lancee, B. and Veit, S. (2021) 'Muslim by Default or Religious Discrimination? Results from a Cross-National Field Experiment on Hiring Discrimination', *Journal of Ethnic and Migration Studies*, **47**, 1305–1326.

Dixon, A. and Telles, E. (2017) 'Skin Color and Colorism: Global Research, Concepts, and Measurement', *Annual Review of Sociology*, **43**, 405–424.

EUROSTAT. (2017) 'First and Second-Generation Immigrants - Statistics on Main Characteristics', *Eurostat Statistics Explained*, accessed at https://ec.europa.eu/eurostat/statistics-explained/index.php?title=First_and_second-generation_immigrants-_statistics_on_main_characteristics#EU_immigrant_population_.E2.80.93_general_overview 5 May 2020.

Farkas, L. (2017) *Data Collection in the Field of Ethnicity*, Luxemburg, Publications Office of the European Union.

Fassmann, H. and Reeger, U. 2012 '"Old" Immigration Countries in Europe'. In Okólski, M. (ed.) *European Immigrations: Trends, Structures and Policy Implications*, Amsterdam, Amsterdam University Press, pp. 65–90.

Fernández-Reino, M., Di Stasio, V. and Veit, S. (2022) 'Discrimination Unveiled: A Field Experiment on the Barriers Faced by Muslim Women in Germany, The Netherlands, and Spain', *European Sociological Review*, doi:10.1093/esr/jcac032

Flores, R. D. (2015) 'The Resurgence of Race in Spain: Perceptions of Discrimination among Immigrants', *Social Forces*, **94**, 237–269.

Flores, R. D. and Telles, E. (2012) 'Social Stratification in Mexico: Disentangling Color, Ethnicity, and Class', *American Sociological Review*, **77**, 486–494.

Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M. and Ambady, N. (2011) 'Looking the Part: Social Status Cues Shape Race Perception', *PLoS One*, **6**, e25107.

Gaddis, S. M. (2017) 'How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies', *Sociological Science*, **4**, 469–489.

Goldberg, D. T. (2006) 'Racial Europeanization', *Ethnic and Racial Studies*, **29**, 331–364.

Goldsmith, A. H., Hamilton, D. and Darity, W. (2006) 'Shades of Discrimination: Skin Tone and Wages', *American Economic Review*, **96**, 242–245.

Gullickson, A. (2005) 'The Significance of Color Declines: A Re-Analysis of Skin Tone Differentials in Post-Civil Rights America', *Social Forces*, **84**, 157–180.

Hersch, J. (2006) 'Skin-Tone Effects among African Americans: Perceptions and Reality', *American Economic Review*, **96**, 251–255.

Koopmans, R., Veit, S. and Yemane, R. (2019) 'Taste or Statistics? A Correspondence Study of Ethnic, Racial and Religious Labour Market Discrimination in Germany', *Ethnic and Racial Studies*, **42**, 233–252.

Lancee, B., Birkelund, G., Coenders, M., Di Stasio, V., Fernández Reino, M., Heath, A., Koopmans, R., Larsen, E., Polavieja, J., Ramos, M., Soiné, H., Thijssen, L., Veit, S. and Yemane, R. (2019a) 'The GEMM Study: A Cross-National Harmonized Field Experiment on Labour Market Discrimination–Codebook', accessed at http://gemm2020.eu/wp-content/uploads/2019/02/GEMM-WP3-codebook.pdf 19 May 2022.

Lancee, B., Birkelund, G., Coenders, M., Di Stasio, V., Fernández-Reino, M., Heath, A., Koopmans, R., Larsen, E., Polavieja, J., Ramos, M., Thijssen, L., Veit, S., Yemane, R. and Zwier, D. (2019b) 'The GEMM Study: A Cross-National Harmonized Field Experiment on Labour Market Discrimination–Technical Report', accessed at http://gemm2020.eu/wp-content/uploads/2019/02/GEMM-WP3-technical-report.pdf 19 May 2022.

Loveman, M. (1999) 'Is "Race" Essential?', *American Sociological Review*, **64**, 891–898.

Macrae, C. N. and Bodenhausen, G. V. (2000) 'Social Cognition: Thinking Categorically', *Annual Review of Psychology*, **51**, 93–120.

Maddox, K. B. (2004) 'Perspectives on Racial Phenotypicality Bias', *Personality and Social Psychology Review*, **8**, 383–401.

Maddox, K. B. and Perry, J. M. (2018) 'Racial Appearance Bias: Improving Evidence-Based Policies to Address Racial Disparities', *Policy Insights from the Behavioral and Brain Sciences*, **5**, 57–65.

Mays, V., Cochran, S. and Barnes, N. (2007) 'Race, Race-Based Discrimination, and Health Outcomes among African Americans', *Annual Review of Psychology*, **58**, 201–225.

Monk, E. (2014) 'Skin Tone Stratification among Black Americans, 2001–2003', *Social Forces*, **92**, 1313–1337.

Monk, E. P. (2016) 'The Consequences of 'Race and Color'', *Social Problems*, **63**, 413–430.

Mörner, M. (1967) *Race Mixture in the History of Latin America*, Boston & London, Little Brown and Company.

National Research Council. (2004) *Measuring Racial Discrimination*, Washington, DC, The National Academic Press.

Pager, D. and Shepherd, H. (2008) 'The Sociology of Racial Discrimination in Employment, Housing, Credit and Consumer Markets', *Annual Review of Sociology*, **34**, 181–209.

Phelps, E. A. and Thomas, L. A. (2003) 'Race, Behaviour and the Brain: The Role of Neuroimaging in Understanding Complex Social Behaviours', *Political Psychology*, **24**, 747–758.

Pierné, G. (2013) 'Hiring Discrimination Based on National Origin and Religious Closeness: Results from a Field Experiment in the Paris Area', *IZA Journal of Labor Economics*, **2**, 1–15.

Polavieja, J. G. (2016) 'Labour-Market Competition, Recession and anti-Immigrant Sentiments in Europe: Occupational and Environmental Drivers of Competitive Threat', *Socio-Economic Review*, **14**, 395–417.

Polavieja, J. G. and Fischer-Souan, M. (2022) 'The Boundary within: Are Applicants of Southern European Descent Discriminated against in Northern European Job Markets?', *Socio-Economic Review*, doi:10.1093/ser/mwac047

Quillian, L., Pager, D., Hexel, O. and Midtbøen, A. H. (2017) 'Meta-Analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring over Time', *Proceedings of the National Academy of Sciences*, **114**, 10870–10875.

Reskin, B. (2012) 'The Race Discrimination System', *Annual Review of Sociology*, **38**, 17–35.

Riach, P. A. and Rich, J. (2002) 'Field Experiments of Discrimination in the Market Place', *The Economic Journal*, **112**, F480–F518.

Richeson, J. A. and Sommers, S. R. (2016) 'Toward a Social Psychology of Race and Race Relations for the Twenty-First Century', *Annual Review of Psychology*, **67**, 439–463.

Roth, W. D. (2016) 'The Multiple Dimensions of Race', *Ethnic and Racial Studies*, **39**, 1310–1338.

Rumbaut, R. G. and Portes, A. (2001) *Ethnicities: Children of Immigrants in America*, Berkeley, Los Angeles, London, University California Press.

Saperstein, A. and Penner, A. M. (2012) 'Racial Fluidity and Inequality in the United States', *American Journal of Sociology*, **11**, 676–727.

Simon, P. (2012) 'Collecting Ethnic Statistics in Europe: A Review', *Ethnic and Racial Studies*, **35**, 1366–1391.

Strabac, Z. and Listhaug, O. (2008) 'Anti-Muslim Prejudice in Europe: A Multilevel Analysis of Survey Data from 30 Countries', *Social Science Research*, **37**, 268–286.

Telles, E. E. and the Project on Ethnicity and Race in Latin America (PERLA). (2014) *Pigmentocracies: Ethnicity, Race and Color in Latin America*, Chapel Hill, NC, University of North Carolina Press.

Thomson, M. and Crul, M. (2007) 'The Second Generation in Europe and the United States: How is the Transatlantic Debate Relevant for Further Research on the European Second Generation?', *Journal of Ethnic and Migration Studies*, **33**, 1025–1041.

Valfort, M. A. (2017) 'La Religion, Facteur de Discrimination à L'embauche en France?', *Revue Économique*, **5**, 167–179.

Veit, S. and Yemane, R. (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.

Weichselbaumer, D. (2017) 'Discrimination against Migrant Job Applicants in Austria: An Experimental Study', *German Economic Review*, **18**, 237–265.

Weiner, M. F. (2014) 'The Ideologically Colonized Metropole: Dutch Racism and Racist Denial', *Sociology Compass*, **8**, 731–744.

Wekker, G. (2016) *White Innocence: Paradoxes of Colonialism and Race*, Durham, Duke University Press.

Wimmer, A. (2013) *Ethnic Boundary Making: Institution, Power and Networks*, New York, Oxford University Press.

Wimmer, A. (2015) 'Race-Centrism: Acritique and a Research Agenda', *Ethnic and Racial Studies*, **38**, 2186–2205.

Wimmer, A. and Soehl, T. (2014) 'Blocked Acculturation: Cultural Heterodoxy among Europe's Immigrants', *American Journal of Sociology*, **120**, 146–186.

Winant, H. (2017) 'Is Racism Global?', *Journal of World-Systems Research*, **23**, 505–510.