Supporting Information (SI Appendix) for

The *Hijab* Penalty:
Feminist Backlash to Muslim Immigrants

Donghyun Danny Choi, Mathias Poertner, Nicholas Sambanis

Contents

A Materials and Methods

A.1 Experimental Design

The experiments focus on exploring whether host population discrimination against immigrants due to intergroup differences in ascriptive characteristics is reduced or eliminated when the immigrant holds progressive, rather than regressive, views with regard to women's role in society. The key outcome variable is the willingness of the host population to offer assistance to immigrants in the context of common day-to-day interactions. The setup and procedures are diagrammatically presented in Figure A1, shown below.

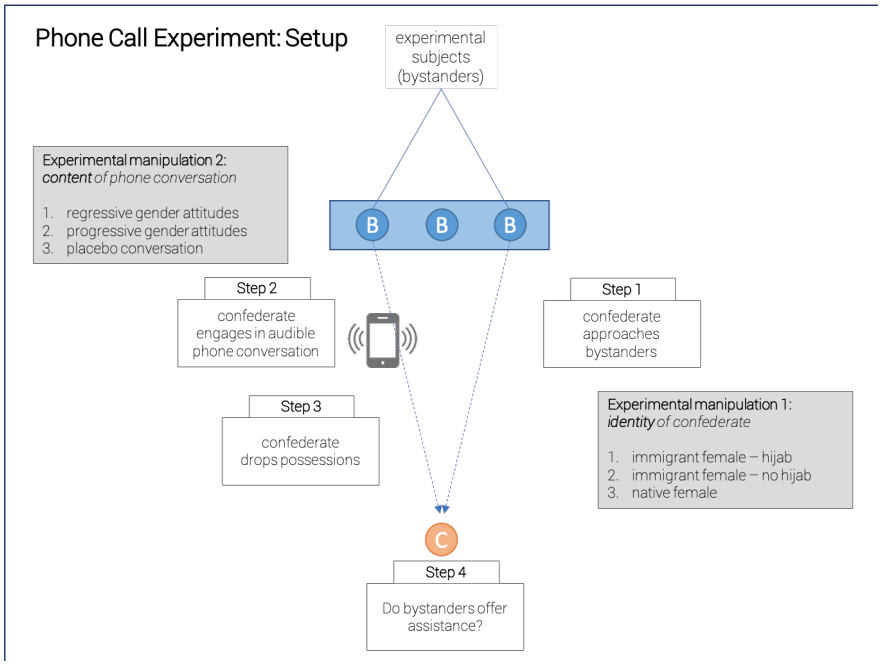


Figure A1: Experimental setup

A.2 Treatment Manipulation

We experimentally manipulated two core dimensions of the intervention.

- **Dimension 1**: Ascriptive characteristics of female confederate conducting the phone call.
    1. Immigrant confederate wearing a hijab
    2. Immigrant confederate wearing plain clothing without hijab
    3. Native confederate (German)

Figure A2: Treatment dimension 1

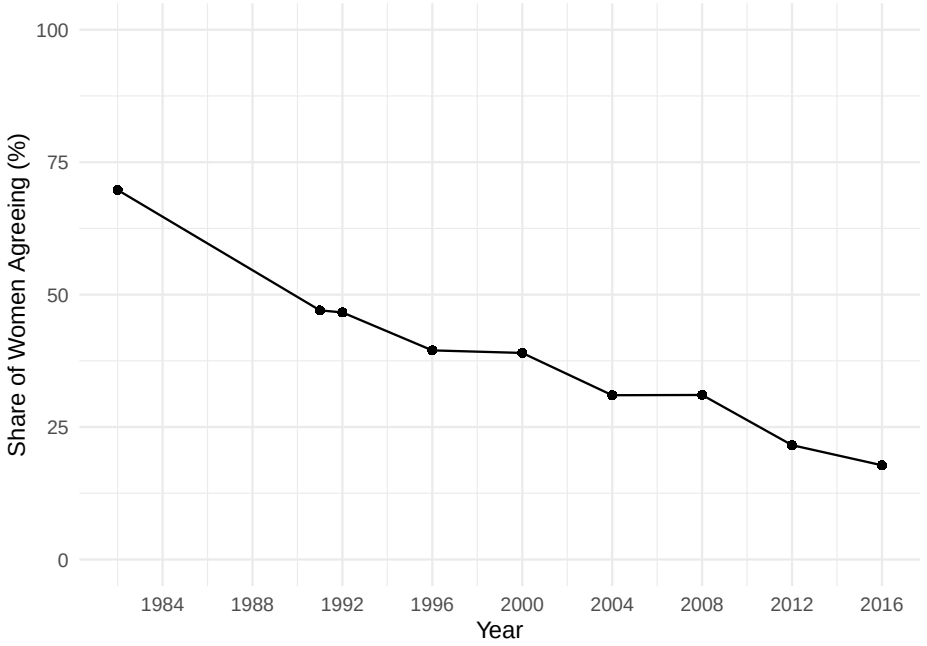
- **Dimension 2**: We also manipulate the content of the phone conversation, to reveal the confederate's attitude towards women's rights. The conversation is intended to be sufficiently loud for bystanders to overhear. The dimension takes on *three* values. (The final sentence in the message indicates status as an immigrant and is omitted in the native confederate conditions.)

    1. **Regressive**: "Hi! Thanks for calling back! I am really mad... My sister is a absent mother [Rabenmutter]. She prefers to work instead of looking after her children and her husband at home. [Pause] I think as a woman she should stay home and look after her family. [*only for immigrant conditions:*] I've never been so mad since we moved to Germany."
    2. **Progressive**: "Hi! Thanks for calling back! I am really happy... I am very proud of my sister. She is pursuing her career; she decided to go to work instead of just looking after her children and her husband at home. [Pause] I think women should not sacrifice their careers just to stay home and look after their family. [*only for immigrant conditions:*] I've never been so happy since we moved to Germany."
    3. **Neutral**: "Hi! Thanks for calling back! Will you come later? [Pause] My sister and I are really looking forward to it. [*only for immigrant conditions:*] I've never been so happy since we moved to Germany."

    Immediately after the last sentence, the confederate drops the lemons and then ends the phone call, saying "Oh, I just dropped something... I will call you back later. Bye."

The specific issue of women's career advancement was chosen because it has been a crucial concern of the women's rights movement in Germany. Although close to 75% of women in Germany agreed with the idea that women should primarily concern themselves with handling the caregiving responsibilities at home while the man is the primary "bread earner" in the early 1980s, there has since been a precipitous drop over time, signaling a generational shift towards equality. By 2016, less than 20 percent of women agreed with the same statement. See Figure A3 for trends over time, as tracked by nationally representative surveys of the German adult population.

Figure A3: Native German Women with Regressive Attitudes about Career Gender Equality



Notes: Share of German women (without immigrant background) who agree completely or rather that "it is much better for everyone involved if the man pursues a professional career and the woman stays at home and looks after the house and children" in nationally representative surveys. Source: GESIS (2020)

A.3 Outcomes

We are interested in measuring the level of assistance offered to the female confederate who drops her possessions (lemons out of a seemingly torn paper bag) in the intervention, as specified in our pre-analysis plan. Enumerators observing each iteration of the intervention collected the following information regarding the reaction of bystanders. Although our unit of analysis is the *iteration*, we collected a mixture of both iteration-level and individual-level outcomes.

- *bystander*: Total number of bystanders within a 3 meter radius of where the iteration is taking place (count)

for each bystander:

- *bystander_gender*: An estimate of each bystander's gender
- *bystander_hp*: Whether each bystander was wearing headphones or earphones (dichotomous)
- *bystander_help*: Whether each bystander offered assistance to the confederate (dichotomous)
- *bystander_age*: An estimate of each bystander's age (dichotomous)

- *bystander_immigrant*: An estimate of whether each has an immigrant background (dichotomous)

Using this information, we construct one main outcome and additional auxiliary outcomes that will be used for the empirical analyses. These outcomes are calculated at the iteration level.

- *help*: Did *any* bystander offer assistance by moving to pick up possessions that the confederate has dropped? (**Calculated at the iteration level.**)

B Logistics and Procedures

B.1 Site Selection

The interventions were conducted at 26 train stations across 25 medium to large-sized cities/towns in the German states of North Rhine-Westphalia (NRW), Saxony, and Lower Saxony. These states were not chosen at random; rather, we arrived at the decision to conduct these interventions in the three states after carefully weighing a combination of state and region-level sociodemographic factors that we believed would be of interest. The most obvious difference between North Rhine-Westphalia (NRW) and Lower Saxony versus Saxony is that they fell under West and East Germany prior to reunification. In addition, these two areas have been traditionally been exposed to very different levels of immigration in Germany's post war history. Whereas NRW and Lower Saxony is considered one of the most ethnically diverse federal states, with the highest proportion of foreign born populations in the country, the two other states have remained relatively ethnically homogeneous. Furthermore, the recent refugee crisis rising as result of the protracted conflict in the Middle East has also had a differential impact on the three states. The Königstein quota system, which combines state level tax revenues and population to assign asylum seekers, has naturally resulted in a high influx of refugees into NRW and Lower Saxony, which also happens to be two of the most populous and affluent states in Germany, and a low influx of refugees to Saxony, which are sparsely populated and lag behind western German states in terms of tax revenue. But perhaps most importantly, there is ample reason to suggest that the level of racial resentment might vary significantly across the west (NRW, Lower Saxony) and the east (Saxony); the level of electoral support for the far-right Alternative für Deutschland (AfD), which primarily campaigned on an anti-immigration agenda, in state and federal elections has been markedly higher in the East in comparison to the west. In some parts of Saxony, the AfD managed to secure the party vote share.

The list of cities and the number of train platforms (in parentheses) at each of the train stations where data collection was implemented is presented below.

- **North Rhine-Westphalia**: Münster (9), Bielefeld (8), Minden (5), Rheine (6), Köln (11), Köln Messe/Deutz (12), Mönchengladbach (9), Neuss (8), Siegen (6), Bonn (5), Düsseldorf (20), Wuppertal (5), Dortmund (31), Duisburg (12), Bochum (8), Gelsenkirchen (6), Hagen (16), Essen (13), Wanne-Eickel (8)
- **Saxony**: Leipzig (21), Görlitz (6), Chemnitz (14), Dresden (16), Zwickau (8)
- **Lower Saxony**: Osnabrück (9), Hannover (12)

B.2 Training

Before the beginning of the intervention in each state, the confederates and enumerators that would observe and code the behavior of the bystanders participated in intensive training workshops led by the authors to ensure a consistently high quality in the delivery of the intervention. These trainings focused on how to select the settings for the intervention, how to play the different roles, how to ensure consistent performances across actors and across teams, and how to code bystander behavior consistently. For the main outcome of the study, whether a bystander provided assistance, enumerators were instructed to code any attempt to offer help in picking up lemons that consisted of a clear physical movement towards the lemons in an effort to help as provision of help, i.e. a clear movement to signal willingness to provide help in picking up lemons was necessary. In order to ensure consistent coding across enumerators and teams, different scenarios were discussed through role-playing activities during the training sessions. These training workshops were followed by extensive test runs in actual train stations with the authors. During the actual data collection, enumerators who were not involved in the intervention observed and coded the bystanders and different enumerators follow up to conduct a post-intervention survey.

We took numerous precautions and trained the confederates and enumerators extensively in procedures to select the sites for the iterations in a way that minimizes the potential for bystanders to witness more than one iteration. First, the specific sites on each train platform were chosen such that it was hard to see the interaction from other platforms (e.g., by making use of walls and signs on the platform, timing the interaction such that stationary trains would block the sight). Second, platforms and the specific sites on those platforms were selected to minimize the chance of repeated participation by the same bystanders. After concluding one iteration on one platform, teams would switch to the platform farthest away from this one that had passengers waiting on it (only train stations with at least four tracks were used). Furthermore, the specific site on that new platform would be chosen to maximize the distance from the previous iteration (e.g., by going to the other end/side). Third, the enumerators tasked with observing the bystanders and coding their behavior were trained to make note of the bystanders for each iteration in order to avoid that—despite the other precautions—bystanders might witness more than one iteration (e.g., if passengers had stayed around after the departure of the train from that platform or had switched platforms). In the very limited instances where the same team conducted interventions at the same train station on more than one day, we conducted field work on different days of the week, choosing a business day and a weekend day in order to minimize chances of commuters being exposed to more than one iteration. Furthermore, enumerators were instructed to begin on the opposite track/side of the train station that during the prior day.

B.3 A Note on Enumerator "Blinding" as to the Purpose of the Project

It was not possible to blind confederates to the general purpose of the experiment. All the coders were intelligent students who were interested in learning about research, thus after a few iterations the coders would have figured out that we were collecting data on bystander behavior across the different treatment conditions. However, we took steps to reduce the risk that coding reflected demand effects and confederates who acted out parts of the scene were expressly told to follow the script and to avoid behaviors that might be designed to elicit specific responses from the bystanders. We did not share the PAP with the actors or coders so they did not know what our prior expectations were

for this experiment. They were given a script to follow during the intervention, were given detailed instructions on how to act, and monitored during the iterations. Finally, there was no normative content in the material we used for the training of confederates (e.g. we referred to measuring assistance to confederates, rather than measuring discrimination and did not use loaded terms such as "bias" or "racism").

B.4 Ethical and Safety Considerations

We took great care to minimize the potential risk to study participants. For a full discussion of these measures, see the research protocol that was reviewed and approved by University of Pennsylvania's Institutional Review Board (IRB Protocol #829824). Beyond our efforts to minimize potential risks to subjects participating in the study, we also took a number of steps to ensure the safety of our research assistants (confederates and enumerators) during the study. Prior to the onset of data collection, we consulted a number of German experts on how to minimize potential risks to our RAs. Furthermore, the other confederates and the enumerators within each team closely monitored the bystanders and stood by, ready to intervene, if necessary, though there was little cause for concern due to the innocuous nature of the phone call and the unobtrusive nature of the intervention. During the training sessions, we discussed potential risks and safety strategies extensively with the research assistants. RAs were instructed to stop the intervention if they felt unsafe at any point. The authors were in constant contact with all teams during the data collection, monitoring their progress and potential safety issues early-on. Last, the German train company, Deutsche Bahn, was instructed about research activities taking place at any given train station on any given day.

B.5 Sampling Protocol for Post-intervention Survey

After each intervention, two enumerators approached the bystanders and conducted a putatively unrelated survey about social life in Germany. The survey instrument is available to readers upon request. Enumerators randomly selected up to two bystanders to interview, following specific instructions regarding sampling. The selection of the interviewees was stratified by their help behavior in order to ensure adequate coverage of helpers and non-helpers in the sample: enumerators chose one bystander who helped and one who did not (or two who did not (two who did), if no one (both) helped). Within each of these two categories, enumerators were instructed to chose the bystander who was closest to the "acting" confederate at the beginning of the iteration. Since the initial location of any given bystander (within each micro-environment) is by design orthogonal to the confederate's initial position and the randomly assigned treatment, this sampling strategy yields a stratified, random sample of bystanders.

C Covariate Balance

To validate whether the random assignment to treatment was successful, we present in Figures A1 and A2 pre-treatment covariate balance across the different comparison across treatment conditions in Figures 3 and 4 of the main paper. While there are a small number of difference-in-means tests that is suggestive of minor imbalance, generally the differences in these pre-treatment covariates are negligible, and we fail to reject F-tests for joint significance.

Table A1: Covariate balance statistics for Figure 3

	# bystanders	# women	# w/earphones	# immigrants	# below 30
Columns (1) and (2)					
Mean Control	2.1393035	1.1475954	0.2857143	0.3333333	1.0000000
Mean Treated	2.0607553	1.0738916	0.2142857	0.2142857	1.5000000
CI Lower	-0.0363452	-0.0324953	-0.2963592	-0.3941554	-1.7160632
CI Upper	0.1934415	0.1799028	0.4392163	0.6322506	0.7160632
P-Value	0.1800768	0.1735746	0.6952922	0.6400586	0.4021699
Columns (2) and (3)					
Mean Control	2.0291734	1.0599676	0.1764706	0.2352941	1.0000000
Mean Treated	2.1393035	1.1475954	0.2857143	0.3333333	1.0000000
CI Lower	-0.2233514	-0.1908154	-0.4949278	-0.5887351	-0.7180895
CI Upper	0.0030913	0.0155599	0.2764404	0.3926566	0.7180895
P-Value	0.0565812	0.0959546	0.5692314	0.6872234	1.0000000
Columns (1) and (3)					
Mean Control	2.0291734	1.0599676	0.1764706	0.2352941	1.0000000
Mean Treated	2.0607553	1.0738916	0.2142857	0.2142857	1.5000000
CI Lower	-0.1433086	-0.1133354	-0.3884269	-0.4017933	-1.6635323
CI Upper	0.0801448	0.0854874	0.3127966	0.4438101	0.6635323
P-Value	0.5792873	0.7835203	0.8269578	0.9196066	0.3777493

Table A2: Covariate balance statistics for Figure 4

	# bystanders	# women	# w/earphones	# immigrants	# below 30
Columns (1) and (3)					
Mean Control	1.9796954	1.0152284	0.4000000	0.4000000	2.8000000
Mean Treated	2.1045455	1.0863636	0.0000000	0.2000000	0.4000000
CI Lower	-0.3181891	-0.2447236	-0.2800874	-0.8996194	-0.2475280
CI Upper	0.0684890	0.1024532	1.0800874	1.2996194	5.0475280
P-Value	0.2050224	0.4209654	0.1778078	0.6707157	0.0676114
Columns (2) and (4)					
Mean Control	2.0937500	1.1041667	0.2500000	0.0000000	1.2500000
Mean Treated	2.1750663	1.1273210	0.2222222	0.3888889	1.0000000
CI Lower	-0.2631763	-0.1713597	-0.7068809	-0.8115027	-2.6698580
CI Upper	0.1005437	0.1563012	0.7624364	0.0337249	3.1698580
P-Value	0.3799406	0.9280487	0.9253973	0.0689661	0.8123129
Columns (2) and (5)					
Mean Control	2.0937500	1.1197917	0.2500000	0.0000000	1.2500000
Mean Treated	2.1565657	1.1414141	0.1111111	0.4444444	1.2222222
CI Lower	-0.2709088	-0.2081274	-0.6040169	-1.0028688	-2.8611886
CI Upper	0.1452775	0.1648824	0.8817947	0.1139799	2.9167442
P-Value	0.5531969	0.8198126	0.6369583	0.1037865	0.9790529
Columns (4) and (6)					
Mean Control	1.9901961	1.0637255	0.0000000	0.0000000	0.8000000
Mean Treated	2.1565657	1.1414141	0.1111111	0.4444444	1.2222222
CI Lower	-0.3639448	-0.2512737	-0.3673338	-1.0028688	-1.4889306
CI Upper	0.0312056	0.0958964	0.1451116	0.1139799	0.6444861
P-Value	0.0986186	0.3794482	0.3465935	0.1037865	0.3902912

D Manipulation Checks

Although the findings presented in the main text of the paper and the appendix suggests that our experimental manipulation was successful, in this section, we provide additional evidence from manipulation checks conducted during our pilot and a partial replication of the intervention with various manipulation checks that the experimental manipulation worked as we had intended.

Table A3: Partial replications with manipulation checks

Outcome	Rate	n
Noticed the confederate	0.996	224
Noticed the call	0.978	224
Recalled treatment direction correctly	0.808	224

More specifically, for our intervention to have been successful, we require that the bystanders 1) noticed our female confederate, 2) noticed that she was engaged in a phone call, and 3) recalled the direction of our treatment in the phone call (progressive vs regressive gender attitudes). The manipulation checks, which we conducted during a pilot in May 2019 and a follow-up study in January 2020, debriefed bystanders who had just been exposed to our experimental intervention, and asked them whether they had in fact noticed both our confederate and the phone call, and whether they could recall whether the phone conversation our confederate had revealed that she had progressive or regressive content with regard to gender (through open-ended questions). The results of this manipulation check exercise is presented in Table A3.

E Iteration Level Analysis

E.1 Full Data

Table A4 shows results on the difference in assistance rates to hijab-wearing immigrants relative to natives at the iteration level, sorted by the content of the phone call. Columns (3) and (4) show that the discrimination against female Muslim immigrants (in column (1)) is driven by assumptions regarding their ideas on women's rights.

Table A5 present results on the effects of the phone call message, comparing progressive to regressive ideas. We find no effect of progressive ideas overall in the full sample (column (1)) or in the native sample (column (4)) and immigrant control condition (3)). However, the positive effects of progressive ideas about gender roles increase assistance rates toward female Muslim immigrants wearing a hijab (column (2)).

Table A4: Hijab versus native comparison: iteration level analysis

	Hijab versus native comparison			
	Any help?			
	(1)	(2)	(3)	(4)
Hijab vs Native	−0.084***	−0.027	−0.134***	−0.091**
	(0.026)	(0.043)	(0.046)	(0.045)
Constant (Control Mean)	0.760***	0.759***	0.761***	0.760***
	(0.018)	(0.029)	(0.033)	(0.033)
Gender Attitude Condition	Pooled	Progressive	Neutral	Regressive
Observations	1,226	418	401	407

[a] Models are estimated with linear regression. Robust standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01.

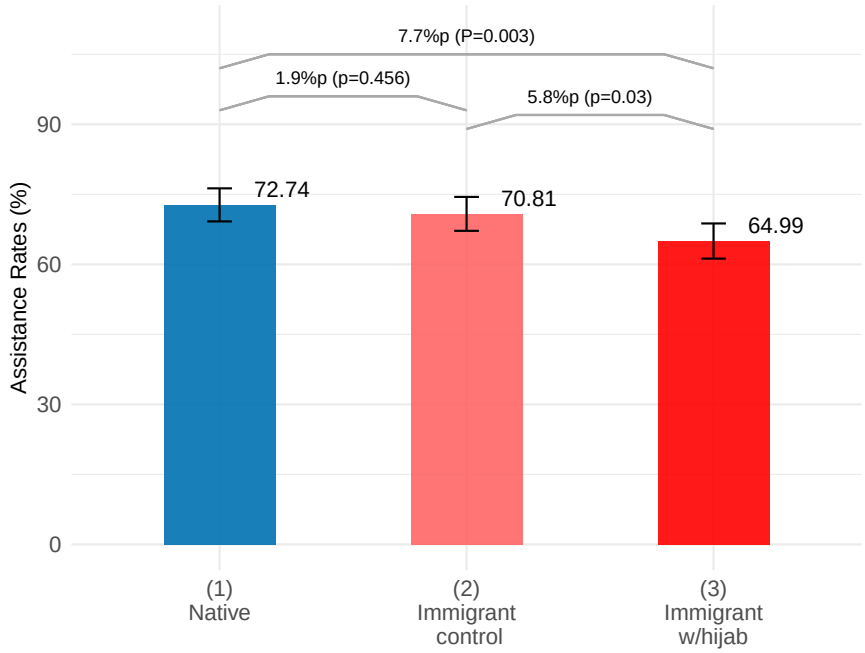
Table A5: Progressive vs regressive gender attitude comparison: iteration level analysis

	Progressive vs regressive message			
	Any help?			
	(1)	(2)	(3)	(4)
Progressive vs Regressive	0.036	0.105**	0.001	−0.002
	(0.026)	(0.046)	(0.046)	(0.042)
Constant (Control Mean)	0.701***	0.627***	0.714***	0.761***
	(0.018)	(0.033)	(0.031)	(0.030)
Confederate Identity Condition	Pooled	Hijab	No Hijab	Native
Observations	1,215	402	396	417

[a] Models are estimated with linear regression. Robust standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01.

E.2 Data Omitting Bystanders Perceived to be Immigrants

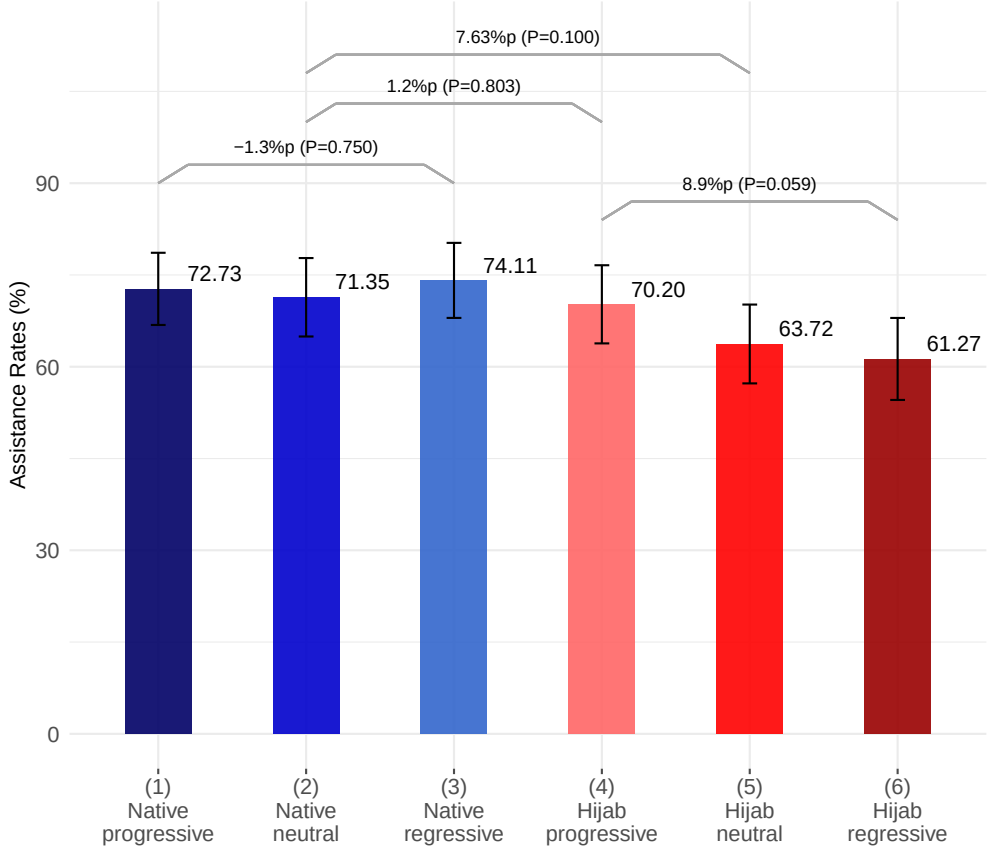
Figure A4: Discrimination against immigrants, bystanders who are perceived as natives



Bars represent the mean rates of assistance for the treatment conditions. The error bars present 95% confidence intervals for the means. The brackets and accompanying information report results of a standard two-tailed difference in means test of treatment conditions with p-values in parentheses.

Figure A4 graphically presents mean rates of assistance for the native, immigrant control, immigrant with hijab conditions among bystanders who are perceived to be native Germans. We omit the behavior of bystanders that were perceived to be of immigrant background by our coders to address that immigrants are affecting our main results. Even with the perceived immigrant bystanders omitted form the analysis, our results regarding discrimination against hijabed immigrants hold (P=0.003).

Figure A5: Offsetting effects of progressive gender attitudes on discrimination, bystanders who are perceived as natives



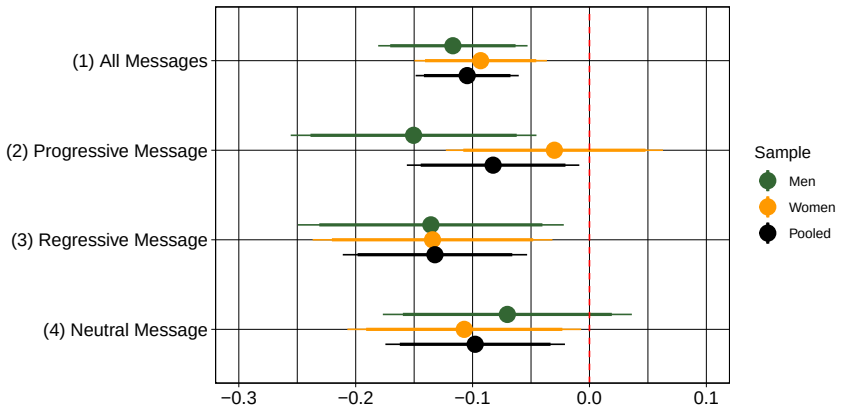
Offsetting effects of progressive gender attitudes on bias. Bars represent the mean rates of assistance for the treatment conditions. The error bars present 95% confidence intervals for the means.The brackets and accompanying information report results of a standard two-tailed difference in means test of treatment conditions with p-values in parentheses.

Figure A5 replicates the offsetting effects of progressive gender attitudes analysis in the main text among bystanders who are perceived to be native Germans. We omit the behavior of bystanders that were perceived to be of immigrant background by our coders to address that immigrants are affecting our main results. The results remain consistent with the analysis reported in the main text.

F Individual Level Analysis

F.1 Full Data

Figure A6: Hijab vs native differences by message type and bystander gender: individual level analysis



The dots represent the point estimate for the hijab versus native comparison, derived from a linear regression model with the same set of fixed effects included in Table A6, columns 1–6. The thin and thick lines represent 95% and 90% confidence intervals. Results for male, female, and pooled bystanders are represented by green, yellow, and black respectively.

Table A6: Effects of Ideas on bias by gender, with number of female bystander fixed effects

	Hijab vs native comparison					
	Outcome: Did an individual bystander help?					
	(1)	(2)	(3)	(4)	(5)	(6)
Hijab vs Native	−0.035	−0.162***	−0.140**	−0.133**	−0.091*	−0.082
	(0.049)	(0.053)	(0.049)	(0.056)	(0.050)	(0.055)
Gender Attitude Condition	Progressive	Progressive	Regressive	Regressive	Neutral	Neutral
Bystander Gender	Female	Male	Female	Male	Female	Male
Fixed Effects	✓	✓	✓	✓	✓	✓
Observations	465	338	415	323	425	326

[a] Models are estimated with linear regression. Robust standard errors clustered at the iteration level in parentheses. *p<0.1; **p<0.05; ***p<0.01.

[b] Fixed effects included number of bystanders at the iteration level, number of *female* bystanders at the iteration level, as well as all individual level attributes that enumerators coded; these included, perceived age bracket, perceived immigrant status, whether or not the bystander was wearing earphones. The number of female bystanders at the iteration level partially assuages concern that women are more susceptible to behavioral spillovers from other female bystanders.

Table A7: Progressive versus regressive attitude comparison by confederate type, disaggregated by gender: individual level analysis

	Progressive versus regressive phone call comparison					
	Did an individual bystander help?					
	(1)	(2)	(3)	(4)	(5)	(6)
Progressive vs Regressive Attitude	0.106**	0.059	−0.028	0.001	−0.015	0.078
	(0.045)	(0.058)	(0.048)	(0.050)	(0.051)	(0.053)
Confederate Identity Condition	Hijab	Hijab	No Hijab	No Hijab	Native	Native
Bystander Gender	Female	Male	Female	Male	Female	Male
Fixed Effects	✓	✓	✓	✓	✓	✓
Observations	441	323	450	339	431	338

[a] Models are estimated with linear regression. Robust standard errors clustered at the iteration level in parentheses. *p<0.1; **p<0.05; ***p<0.01.

[b] Fixed effects included number of bystanders at the iteration level, as well as individual level attributes including perceived age bracket, and whether or not the bystander was wearing earphones.

Table A8: Progressive versus regressive attitude comparison by confederate type, disaggregated by gender: individual level analysis, including number of female bystander fixed effects

	Progressive versus regressive phone call comparison					
	Did an individual bystander help?					
	(1)	(2)	(3)	(4)	(5)	(6)
Progressive vs Regressive Attitude	0.096**	0.063	−0.022	0.025	−0.029	0.082
	(0.047)	(0.060)	(0.050)	(0.053)	(0.051)	(0.054)
Confederate Identity Condition	Hijab	Hijab	No Hijab	No Hijab	Native	Native
Bystander Gender	Female	Male	Female	Male	Female	Male
Fixed Effects	✓	✓	✓	✓	✓	✓
Observations	423	316	432	323	431	329

[a] Models are estimated with linear regression. Robust standard errors clustered at the iteration level in parentheses. *p<0.1; **p<0.05; ***p<0.01.

[b] Fixed effects included number of bystanders at the iteration level, number of *female* bystanders at the iteration level, as well as all individual level attributes that enumerators coded; these included perceived age bracket, perceived immigrant status, whether or not the bystander was wearing earphones. The number of female bystanders at the iteration level partially assuages concern that women are more susceptible to behavioral spillovers from other female bystanders.

F.2 Data Omitting Bystanders Perceived to be Immigrants

Table A9: Effects of ideas on bias by gender, perceived native German bystanders

	Hijab vs native comparison					
	Outcome: Did an individual bystander help?					
	(1)	(2)	(3)	(4)	(5)	(6)
Hijab vs Native	−0.031	−0.164***	−0.132**	−0.144**	−0.102**	−0.083
	(0.050)	(0.054)	(0.050)	(0.056)	(0.050)	(0.055)
Gender Attitude Condition	Progressive	Progressive	Regressive	Regressive	Neutral	Neutral
Bystander Gender	Female	Male	Female	Male	Female	Male
Fixed Effects	✓	✓	✓	✓	✓	✓
Observations	449	320	407	315	418	316

[a] Models are estimated with linear regression. Robust standard errors clustered at the iteration level in parentheses. *p<0.1; **p<0.05; ***p<0.01.

[b] Fixed effects included number of bystanders at the iteration level, number of *female* bystanders at the iteration level, as well as individual level attributes that enumerators coded; these included, perceived age bracket, whether or not the bystander was wearing earphones. The number of female bystanders at the iteration level partially assuages concern that women are more susceptible to behavioral spillovers from other female bystanders.

Table A10: Progressive versus regressive attitude comparison by confederate type, disaggregated by gender: individual level analysis

	Progressive versus regressive phone call comparison					
	Did an individual bystander help?					
	(1)	(2)	(3)	(4)	(5)	(6)
Progressive	0.102**	0.067	−0.026	−0.013	−0.008	0.080
vs Regressive Attitude	(0.047)	(0.058)	(0.048)	(0.052)	(0.052)	(0.053)
Confederate Identity Condition	Hijab	Hijab	No Hijab	No Hijab	Native	Native
Bystander Gender	Female	Male	Female	Male	Female	Male
Fixed Effects	✓	✓	✓	✓	✓	✓
Observations	426	311	434	328	430	324

[a] Models are estimated with linear regression. Robust standard errors clustered at the iteration level in parentheses. *p<0.1; **p<0.05; ***p<0.01.

[b] Fixed effects included number of bystanders at the iteration level, as well as all individual level attributes that enumerators coded; these included perceived age bracket, whether or not the bystander was wearing earphones.

G Conditional Effects (Post-Treatment Survey)

Table A11: Effect of the progressive gender attitudes, disaggregated by bystander religion: post intervention survey sample

	Progressive versus regressive message					
	Did an individual offer help?					
	(1)	(2)	(3)	(4)	(5)	(6)
Progressive vs Regressive, Hijab (H6A)	0.178***	0.160**	−0.004	−0.009	0.240***	0.215**
	(0.066)	(0.068)	(0.152)	(0.160)	(0.088)	(0.092)
Sample	Full Sample	Full Sample	Christian	Christian	Not Religious	Not Religious
# of Bystander FE	Yes	Yes	Yes	Yes	Yes	Yes
Bystander Attribute FE	No	Yes	No	Yes	No	Yes
Observations	230	220	53	49	109	105
R^2	0.176	0.191	0.170	0.207	0.247	0.245

Notes: Models are estimated with linear regression. Robust standard errors clustered at the iteration level in parentheses. *p<0.1; **p<0.05; ***p<0.01. Columns (3) and (4) subset to individuals who self-identified as either Christian in the post-intervention survey (protestant and catholic). Columns (5) and (6) subset to individuals who self-reported as having "no religion." Bystander attribute fixed effects includes all individual level attributes that enumerators coded; perceived age bracket, perceived immigrant status, whether or not the bystander was wearing earphones.

Drawing on data from the post-intervention survey, we can take a closer look at the effect of religious identity and education levels on shaping attitudes toward Muslim immigrants as a function of the gender-specific ideological message conveyed in the phone call experiment. Table A11 shows that the progressive message increases help to hijab-wearing Muslims (column 2, 16.0%p) while controlling for the number of bystanders as well as bystander-attribute fixed effects (e.g. wearing ear phones). This effect is much larger for bystanders who declare no religion (column 6, 21.5%p) than for those who report that they are religious Christians (column 4, -0.9%p). Due to high attrition rates in the survey, we are limited in the analyses of conditional effects we can do. However, the results indicate that the progressive gender roles message resonates with secular bystanders, consistent with our theoretical expectations.

Table A12: Effect of the progressive gender attitudes, disaggregated by bystander religion: post intervention survey sample, weighted by proportion of helpers and non-helpers in the experimental sample

	Progressive versus regressive message					
	Did an individual offer help?					
	(1)	(2)	(3)	(4)	(5)	(6)
Progressive vs Regressive, Hijab (H6A)	0.120*	0.102	−0.063	−0.066	0.192**	0.168*
	(0.067)	(0.068)	(0.149)	(0.157)	(0.090)	(0.093)
Sample	Full Sample	Full Sample	Christian	Christian	Not Religious	Not Religious
# of Bystander FE	Yes	Yes	Yes	Yes	Yes	Yes
Bystander Attribute FE	No	Yes	No	Yes	No	Yes
Observations	230	220	53	49	109	105

Notes: Models are estimated with linear regression. Robust standard errors clustered at the iteration level in parentheses. *p<0.1; **p<0.05; ***p<0.01. Columns (3) and (4) subset to individuals who self-identified as either Christian in the post-intervention survey (protestant and catholic). Columns (5) and (6) subset to individuals who self-reported as having "no religion." Bystander attribute fixed effects includes all individual level attributes that enumerators coded; perceived age bracket, perceived immigrant status, whether or not the bystander was wearing earphones.

We also present results of the same analysis presented in Table A11, weighted by the proportion of helpers and non-helpers in the experimental sample for H6A (progressive vs regressive, hijab). The findings are reported in Table A12. Although the treatment effects for the full sample are somewhat diminished, we still find strong effects among non-religious people in the sample, as reported in columns (5) and (6). We interpret these findings to be in line with the results reported in Table A11.

Table A13: Effect of the progressive gender attitudes, disaggregated by bystander religion: post intervention survey sample

	Progressive versus regressive message					
	Did an individual offer help?					
	(1)	(2)	(3)	(4)	(5)	(6)
Progressive vs Regressive, Hijab (H6A)	0.303	0.349*	−0.211	−0.237	0.303	0.338
	(0.220)	(0.204)	(0.185)	(0.184)	(0.220)	(0.213)
Atheist	0.081	−0.002	−0.254	−0.336**	0.081	−0.005
	(0.177)	(0.192)	(0.164)	(0.168)	(0.177)	(0.202)
Female	0.236	0.206				
	(0.180)	(0.192)				
H6A × Atheist	−0.161	−0.222	0.495**	0.528**	−0.161	−0.202
	(0.258)	(0.244)	(0.222)	(0.217)	(0.258)	(0.258)
H6A × Female	−0.514*	−0.583**				
	(0.264)	(0.249)				
Atheist × Female	−0.335	−0.328				
	(0.231)	(0.235)				
H6A × Atheist × Female	0.656**	0.741**				
	(0.324)	(0.308)				
Confederate Identity Condition	Hijab	Hijab	Hijab	Hijab	Hijab	Hijab
Sample	Full	Full	Female	Female	Male	Male
# of Bystander FE	Yes	Yes	Yes	Yes	Yes	Yes
Bystander Attribute FE	No	Yes	No	Yes	No	Yes
Observations	162	154	86	82	76	72

[a] Models are estimated with linear regression. Robust standard errors clustered at the iteration level in parentheses. *p<0.1; **p<0.05; ***p<0.01, one-tailed test.

[b] Columns (3) and (4) subset to female bystanders. Columns (5) and (6) subset to male bystanders.

[c] Bystander attribute fixed effects includes all individual level attributes that enumerators coded; perceived age bracket, perceived immigrant status, whether or not the bystander was wearing earphones.

We anticipated that the treatment effects of the progressive versus regressive message would likely be driven by female bystanders who themselves hold a progressive outlook with regard to womens' role in society. We further expected that non-religious (atheist) women would be much more likely to hold progressive views, and thus respond to the progressive message more than other subgroups of the population. In Table A13, we conduct regression analysis of individual level treatment effects interacted by whether the bystander who completed the post-intervention survey self-identified as non-religious and was a female. Our results indicate that women who are non-religious are indeed more responsive to our progressive message treatment. The final row of columns (1) and (2), which utilizes the full survey response (male and female) data, shows a significant and positive interaction effect, suggesting that our posited mechanism is likely to be valid. These results are replicated in columns (3) and (4), where we just subset to female bystander-survey respondents.

Table A14: Effect of the progressive gender attitudes, disaggregated by bystander religion: post intervention survey sample, weighted by proportion of helpers and non-helpers in the experimental sample

	Progressive versus regressive message					
	Did an individual offer help?					
	(1)	(2)	(3)	(4)	(5)	(6)
Progressive vs Regressive, Hijab (H6A)	0.238	0.287	−0.274	−0.292	0.238	0.276
	(0.225)	(0.206)	(0.178)	(0.177)	(0.225)	(0.216)
Atheist	0.084	−0.009	−0.253	−0.343**	0.084	−0.011
	(0.185)	(0.195)	(0.160)	(0.159)	(0.185)	(0.204)
Female	0.234	0.198				
	(0.182)	(0.186)				
H6A × Atheist	−0.164	−0.214	0.493**	0.534**	−0.164	−0.195
	(0.263)	(0.245)	(0.219)	(0.212)	(0.263)	(0.260)
H6A × Female	−0.512*	−0.577**				
	(0.265)	(0.245)				
Atheist × Female	−0.336	−0.330				
	(0.234)	(0.229)				
H6A × Atheist × Female	0.657**	0.742**				
	(0.325)	(0.303)				
Sample	Full	Full	Female	Female	Male	Male
# of Bystander FE	Yes	Yes	Yes	Yes	Yes	Yes
Bystander Attribute FE	No	Yes	No	Yes	No	Yes
Observations	162	154	86	82	76	72

[a] Models are estimated with linear regression. Robust standard errors clustered at the iteration level in parentheses. *p<0.1; **p<0.05; ***p<0.01, one-tailed test.

[b] Columns (3) and (4) subset to female bystanders. Columns (5) and (6) subset to male bystanders.

[c] Bystander attribute fixed effects includes all individual level attributes that enumerators coded; perceived age bracket, perceived immigrant status, whether or not the bystander was wearing earphones.

We also present results of the same analysis presented in Table A13, weighted by the proportion of helpers and non-helpers in the experimental sample for H6A (progressive vs regressive, hijab). The findings are reported in Table A14. They do not substantively change the results reported in Table A13.

Table A15: Effect of the progressive gender attitudes, disaggregated by bystander education: post intervention survey sample

	Progressive versus regressive message		
	Did an individual offer help?		
	(1)	(2)	(3)
Progressive versus Regressive, Hijab (H6A)	−0.022	−0.022	0.265**
	(0.131)	(0.131)	(0.113)
High Education	0.029		
	(0.131)		
H6A × High Education	0.287*		
	(0.169)		
Constant	0.429***	0.429***	0.457***
	(0.098)	(0.098)	(0.086)
Sample Gender	Female	Female	Female
Sample Education	Full	Low	High
Observations	131	60	71
R^2	0.069	0.001	0.073
Note:			*p<0.1; **p<0.05; ***p<0.01

In order to further validate our intuition regarding the treatment effect being driven by *female* bystanders who hold a progressive outlook with regard to womens' role in society, we look at heterogeneity in the treatment effect for the progressive versus regressive message based on the level of education, a strong correlate of gender attitudes in the German context. The results are presented in Table A15. For observations in the post-intervention survey, we collected information on the level of education for the bystanders, and created a dummy variable "high education" to denote individuals who passed the university entrance qualification exam (Abitur) or obtained bachelors, masters, or doctoral degrees. We interacted (and also subsetted) this dummy with the treatment indicator for the progressive vs regressive message for hijab confederates (Column 1). We find that, as expected, women who are highly educated (and thus more likely to hold progressive views on gender) are significantly more likely to be responsive to the progressive vs regressive message treatment than those who are not.

Table A16: Effect of the progressive gender attitudes, disaggregated by bystander education: post intervention survey sample, weighted by proportion of helpers and non-helpers in the experimental sample

	Progressive versus regressive message		
	Did an individual offer help?		
	(1)	(2)	(3)
Progressive versus Regressive, Hijab (H6A)	−0.090	−0.090	0.196*
	(0.133)	(0.133)	(0.113)
High Education	0.029		
	(0.132)		
H6A × High Education	0.286*		
	(0.170)		
Constant	0.501***	0.501***	0.530***
	(0.100)	(0.100)	(0.086)
Sample Gender	Female	Female	Female
Sample Education	Full	Low	High
Observations	131	60	71
Note:			*p<0.1; **p<0.05; ***p<0.01

We also present results of the same analysis presented in Table A15, weighted by the proportion of helpers and non-helpers in the experimental sample for H6A (progressive vs regressive, hijab). The findings are reported in Table A16. They do not substantively change the results reported in Table A15.

Table A17: Lack of evidence on differential response/attrition in the post-treatment survey

	Treated Mean	Control Mean	Diff. in Means	T-Test P-Value
Progressive vs Regressive Gender Attitude Comparison, Hijab Confederates Only				
Female	0.5826087	0.5619048	0.0207039	0.7578991
Atheist	0.4782609	0.4761905	0.0020704	0.9756417
Christian	0.2173913	0.2285714	-0.0111801	0.8432199
Earphones	0.0347826	0.0380952	-0.0033126	0.8964921
Native	1.0000000	0.9809524	0.0190476	0.1582902
Joint F-Statistic: 0.5014	**P-Value=0.7750**			

Although we adjust the regression analysis in Table A13 with number of bystander fixed effects as well as bystander attribute fixed effects, some might still be concerned that the post-intervention survey is susceptible to differential attrition in responses across treatment conditions. If this were the case, any findings that utilizes data collected through the post-intervention survey might be driven by differences in the characteristics of individuals by treatment condition. In order to assuage this concern, we present results from a simple difference in means test on the survey respondent's characteristics across the progressive and regressive gender attitude conditions presented in Table A13. The results are prsented in Table A17. Across all covariates that can likely be considered pre-treatment, we have very strong balance across the progressive message vs regressive message conditions; the t-test for each of the 5 covariates fail to reach statistical significance at conventional levels, and the magnitude of the differences are small. The joint F-statistic is also insigifnicant, with a p-value of 0.822. These results should alleviate much of the concern that the findings in Table A13 are merely reflective of systematic differences in bystander characteristics among people who answered the post-intervention survey.

H Potential Behavioral Spillovers

Table A18: Help rates by bystander gender composition

	Help Rates					
	Number of Women Bystanders in Iteration?					
	n(women)=5	n(women)=4	n(women)=3	n(women)=2	n(women)=1	n(women)=0
Help Rate n(bystander)=5	0.400 (0.167)	0.347 (0.057)	0.200 (0.055)	0.200 (0.047)	0.100 (0.100)	0.500 (0.208)
Observations	25	75	60	55	10	20
Help Rate n(bystander)=4		0.375 (0.076)	0.265 (0.034)	0.357 (0.035)	0.250 (0.081)	0.333 (0.083)
Observations		80	200	244	48	12
Help Rate n(bystander)=3			0.458 (0.041)	0.377 (0.027)	0.432 (0.028)	0.441 (0.057)
Observations			192	435	345	93
Help Rate n(bystander)=2				0.508 (0.025)	0.534 (0.018)	0.624 (0.032)
Observations				478	914	234
Help Rate n(bystander)=1					0.680 (0.023)	0.739 (0.022)
Observations					419	395

When conducting individual level analyses of helping behavior, it may be of concern to some that bystanders might adjust their behavior in accordance with the behavior of others. We partially address the potential for these "behavioral spillovers" in the regression analyses in Table A6 and A8 by including the number of bystander fixed effects, as well as the number of female bystander fixed effects which should allay some concern over how the size and gender composition of the bystander pool affects our individual level helping behavior. However, in Table A18, we also present the mean individual-level assistance rate by the size and gender composition of the bystander pool at the iteration level. In general, we observe that as the number of bystanders (or size of the bystander pool) at the iteration level increases, the mean rate of assistance decreases; this is consistent with the notion that a sole bystander might feel highly pressured to help our confederate in collecting her possession since there is no one else near by who is offering assistance. The decrease in the assistance rates seem to be relatively monotonic, as seen in the *cross-row* comparisons. Second, we also observe that there are no clear patterns of heterogeneity in individual assistance rates based on the gender composition of the bystander pool, as seen in the *cross-column* comparisons. These observations taken together suggest that behavioral spillovers are unlikely to pose a huge threat to individual-level estimates of our experimental treatment effects, and should partially be remedied by the fixed effects approach taken in Tables A6 and A8.

Table A19: Gender Spillovers (Iteration Level)

	Help Rates	
	Male Bystander Help	Woman Bystander Help
	(1)	(2)
Hijab vs Native (H2A)	−0.062	−0.016
	(0.090)	(0.072)
Woman Bystander Present	−0.456***	
	(0.070)	
H2A × Woman Present	0.011	
	(0.103)	
Man Bystander Present		−0.471***
		(0.063)
H2A × Man Present		0.118
		(0.092)
Constant	0.786***	0.733***
	(0.061)	(0.049)
Observations	418	418
Note:		*p<0.1; **p<0.05; ***p<0.01

In Table A19, we address the concern that differential help rates to hijabed (Muslim) women might be driven by male bystanders who are unsure if their helpful intervention would be welcomed by Muslim women. If this logic were to hold, we would expect that male bystanders would hold off on helping a Muslim woman in the presence of female bystanders, as they feel that women would be less threatening.

In order to probe this intuition, we created a set of alternative outcomes that codes (at the iteration level) whether any male bystander offered assistance and whether any female bystander offered assistance. We also created a dummy variable that takes on a value of "1" when there is a woman bystander present at the scene. We run a regression in which we interact this dummy variable with the treatment indicator for hijab vs native comparisons. In column (1) we examine whether the presence of a woman bystander affects the helping behavior of male bystanders with respect to treatment. Although we observe that men are less likely to assist both hijabed and native women when there is a woman bystander present *overall*—already made clear in Table A18—there is no evidence of heterogeneous effects; the presence of the woman bystander does not moderate the differential help rates between hijabed (Muslim) vs native confederates. The same applies for female bystander behavior in the presence of male bystanders, presented in column (2).

I Effects Disaggregated by Former East vs West Germany

In response to reviewer suggestions to examine whether political context conditions the treatment effects in our field experiment, we disaggregate the two main findings—a) discrimination against hijab-wearing immigrants, and b) the offsetting effect of the progressive gender attitude—by whether the iterations were conducted in states that fell in either Former West and East Germany. Our intuition behind why we expect that there might be some heterogeneity across the Former West and East are discussed in some detail in section 2.1. We expected that given the electoral support for the AfD in state and local elections in the Former East, that discrimination effects against immigrant minorities would be significantly larger in the East than in the West.

Table A20: Discrimination against Hijab Immigrants, Former West/East Germany

	Hijab versus native							
	Any help?							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Hijab vs Native (H2), Pooled	−0.070**	−0.116***						
	(0.032)	(0.043)						
H2, Progressive Message			−0.044	0.009				
			(0.052)	(0.074)				
H2, Neutral Message					−0.121**	−0.142**		
					(0.059)	(0.068)		
H2, Regressive Message							−0.043	−0.198**
							(0.054)	(0.081)
Constant	0.731***	0.823***	0.750***	0.781***	0.702***	0.863***	0.737***	0.818***
	(0.022)	(0.030)	(0.036)	(0.052)	(0.043)	(0.047)	(0.039)	(0.060)
Region	West	East	West	East	West	East	West	East
Observations	833	393	292	126	260	141	281	126
R^2	0.006	0.019	0.002	0.0001	0.016	0.031	0.002	0.047
Note:								*p<0.1; **p<0.05; ***p<0.01

We present the tests for the hijab vs native comparison reported in Figure 4 of the main text in tabular form, in Table A20. As predicted, we observe that the ATE estimate for the hijab vs native comparison is larger in the iterations run in the Former East (11.6%p) than those run in the Former West (7%p), although the differences between the treatment effects are not statistically distinguishable. Some interesting patterns emerge when we disaggregate by the content of the message—progressive, neutral, regressive. The native hijab comparisons in which confederates signalled progressive gender attitudes are statistically indistinguishable from zero, meaning that once the hijab immigrant signaled their progressive outlook with regard to women, discrimination against them decreased in both the East and West. Differences between native and hijab conditions persist in most of the message conditions.

Although our intuition regarding the treatment effects for the progressive vs regressive gender attitudes are less clear, we nonetheless disaggregate the effect of gender attitudes by whether iterations were run in the Former West versus the East. Results are reprted in Table A21. Note that the offsetting

Table A21: Progressive vs Regressive Message Effects, Former West/East Germany

	Progressive vs Regressive Message							
	Any help?							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Prog vs Reg (H6), Pooled	0.057*	−0.001						
	(0.032)	(0.044)						
H6, Hijab			0.125**	0.070				
			(0.058)	(0.076)				
H6, No Hijab					−0.003	0.010		
					(0.054)	(0.087)		
H6, Native							0.048	−0.082
							(0.054)	(0.065)
Constant	0.667***	0.768***	0.581***	0.721***	0.716***	0.710***	0.702***	0.863***
	(0.023)	(0.030)	(0.041)	(0.052)	(0.036)	(0.058)	(0.040)	(0.045)
Region	West	East	West	East	West	East	West	East
Observations	836	379	272	130	284	112	280	137
R^2	0.004	0.00000	0.017	0.007	0.00001	0.0001	0.003	0.012
Note:								*p<0.1; **p<0.05; ***p<0.01

effect of the progressive vs regressive gender attitude we report in Figure 5 of the main text was for the iterations with immigrant confederates donning a hijab. We find that when we disaggregate the effects by Former West vs East, the effects are larger in the West by around 5% points (12.5%p vs 7.0%p).

Overall, we find suggestive evidence that political/geographic context may condition the treatment effects. Exploring regional differences can be helpful in adjudicating among rival mechanisms underlying our results. For example, one might conjecture that the progressive treatment exposes bystanders not only to ideas about gender norms, but also about the confederate's work ethic. A "work ethic" interpretation of our treatment would be inconsistent with the gender differences in outcomes we have observed (we would have expected equally strong effects among men under such an interpretation of the treatment), and we believe that women who choose to work at home could also have a strong work ethic. Nonetheless, we can explore regional differences in outcomes to think more about this question. To the extent that female labor market participation was historically larger in the East due to legacies of the Communist system, these results also suggest that differences in work ethic or participation in the formal economy cannot drive the results that we report in the paper (since such differences would have suggested larger effects in the East vs the West). There are other differences between East and West, including significant differences in the density of immigrant populations, which results in more frequent and more varied forms of inter-group contact between natives and immigrants in the West. These differences could be relevant to our findings. While we cannot make definitive causal claims regarding the East vs West differences, we believe that these analyses opens up the opportunity for future work to probe precisely why these difference may be observed.

References

GESIS. 2020. "Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS - Kumulation 1980-2018." Leibniz-Institut für Sozialwissenschaften, GESIS Datenarchiv, Köln. ZA5274 Datenfile Version 1.0.0, https://doi.org/10.4232/1.13395.