Online Appendix for

Temperature and Outgroup Discrimination

1 Materials and Methods

Experimental design

The experiments focus on exploring whether host population discrimination against immigrants due to intergroup differences in ascriptive characteristics is reduced or eliminated by immigrants' linguistic assimilation. 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 S1, shown below.

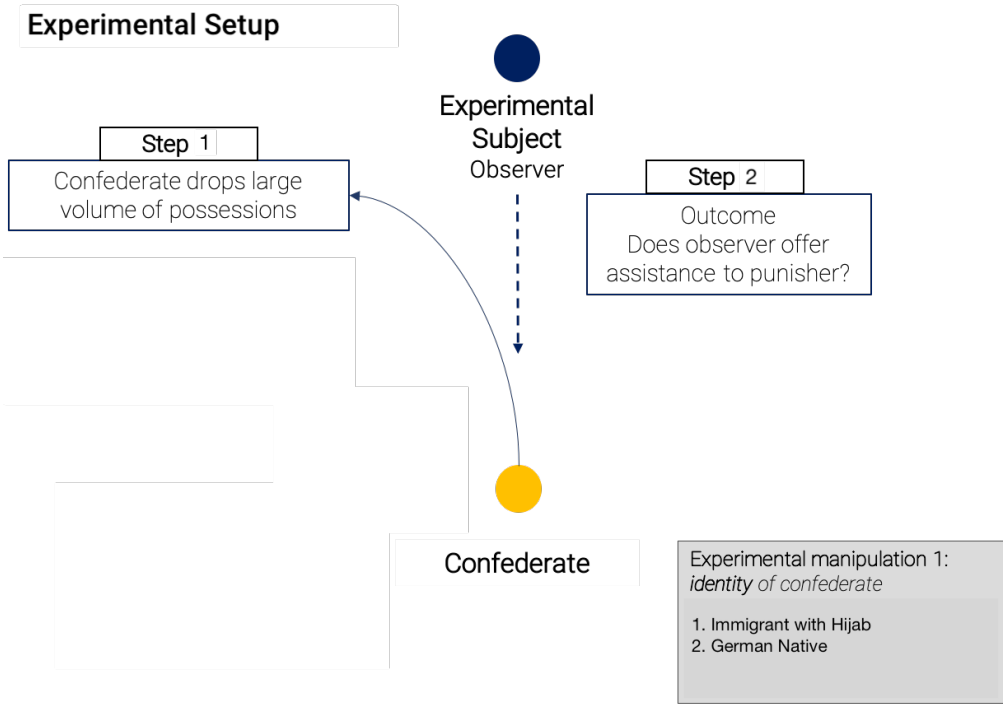

Figure S1: Experimental Setup

Treatment manipulation

For this study, we experimentally manipulated ascriptive characteristics of female confederate:

1. Immigrant confederate wearing a hijab
2. Native German confederate

Figure S2: Treatments

Temperature Measures

Immediately prior the execution of any given iteration, the enumerators took temperature measurements on the specific train platform using a digital precision thermometer and recorded these measurements along with their observations of help behavior by the bystanders.

Outcomes

We are interested in measuring the level of assistance offered to the female confederate who drops her possessions (bag of oranges or lemons) across a broad range of ambient temperature points, as specified in our pre-analysis plan. Enumerators observing each iteration of the intervention collected the following information regarding the reaction of bystanders. This information was collected at the level of the iteration, which constitutes our unit of analysis.

- *bystander*: Total number of bystanders within a 3 meter radius of where the iteration is taking place (count)
- *bystanderHP*: Total number of bystanders with headphones or earphones (count)
- *help*: Whether any bystander offered assistance to the confederate (dichotomous)

Using this information, we construct the outcome that will be used for the empirical analyses (calculated at the iteration level):

- *help*: Did *any* bystander offer assistance by moving to pick up possessions that the confederate dropped?

2 Logistics and Procedures

Site selection

The interventions were conducted at 30 train stations across 29 medium to large-sized cities and towns in the German states of North Rhine-Westphalia (NRW), Brandenburg, Saxony, and Lower Saxony. These states were not chosen at random; rather, we arrived at the decision to conduct these interventions in the four 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 NRW and Lower Saxony versus the two other states (Brandenburg, and Saxony) is that they were part of West and East Germany respectively 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 is considered one of the most ethnically diverse federal states, Brandenburg and Saxony 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 four 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 Lower Saxony and NRW, which also happens to be one of the most populous and affluent states in Germany, and a low influx of refugees to Brandenburg and 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, Brandenburg); 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 largest 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 (Hbf) (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)
- **Saxony**: Leipzig (21), Görlitz (6), Chemnitz (14), Dresden (16), Zwickau (8)
- **Lower Saxony**: Osnabrück (9), Hannover (12)
- **Brandenburg**: Potsdam (7), Forst (5), Cottbus (10), Frankfurt-Oder (12), Brandenburg (6)

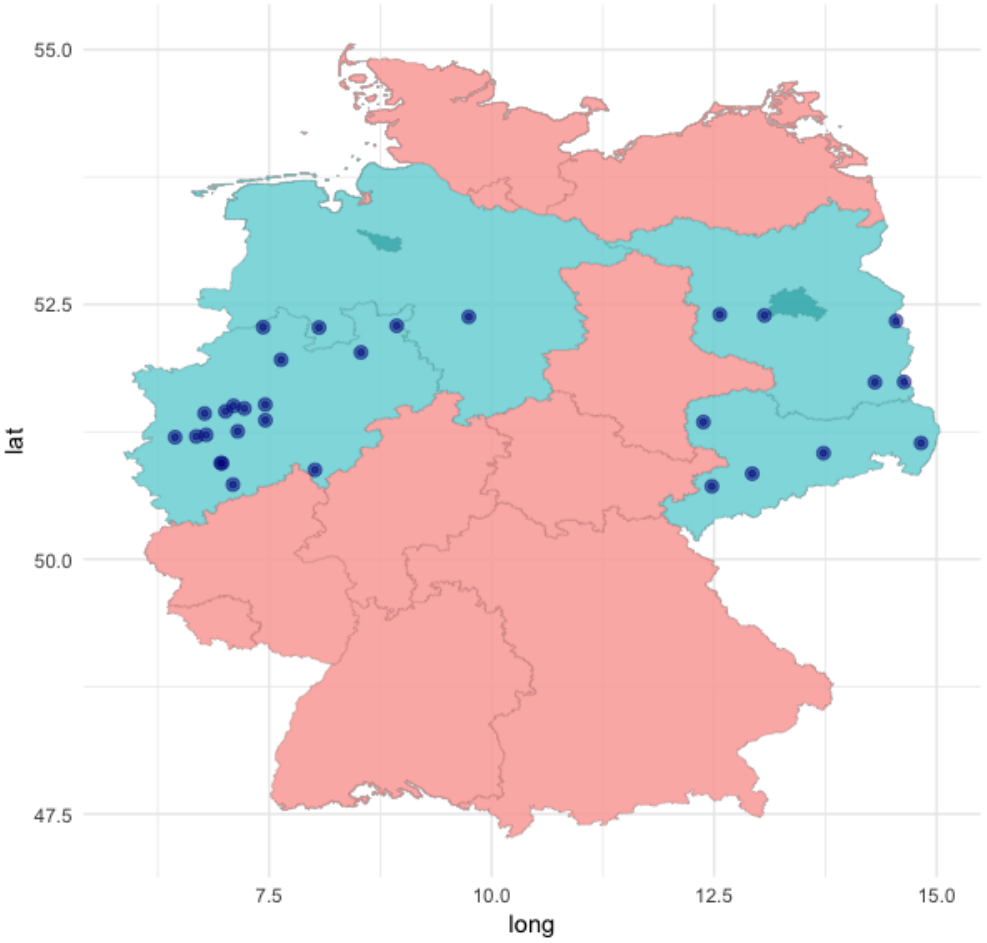

Figure S3: Study sites – 30 train stations across 4 states in North Rhine-Westphalia, Saxony, Lower Saxony, and Brandenburg

Figure S4: Example of a Typical Train Platform Section (at Bielefeld Central Station)

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 day-long 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 outcome of the study, whether a bystander provided assistance, enumerators were instructed to code any attempt to offer help in picking up oranges/lemons that consisted of a clear physical movement towards the oranges in an effort to help as provision of help, i.e. a clear movement to signal willingness to provide help in picking up oranges/lemons was necessary. In order to ensure consistent coding across enumerators and teams, different scenarios were practiced and 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.

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 five 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 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 previous day.

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").

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 the University of Pennsylvania Institutional Review Board (IRB Protocols #829824 and #833206); a waiver of the consent process was obtained.) 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.

3 Additional Analyses

Table S1: Descriptive Statistics

Variable	Min	Max	Mean	Median	SD	N
Temperature	16.1	41.4	27.13	26.8	4.38	1786
Assistance (Outcome)	0.0	1.0	0.72	1.0	0.45	1786
Share of women bystanders	0.0	1.0	0.54	0.5	0.35	1786
Share of bystanders with earphones	0.0	1.0	0.06	0.0	0.17	1785
Share native bystanders	0.0	1.0	0.94	1.0	0.21	1121
Share of bystanders below 30	0.0	1.0	0.32	0.0	0.42	1121
Share of bystanders above 60	0.0	1.0	0.20	0.0	0.36	1121
Share of Christian bystanders	0.0	1.0	0.37	0.0	0.46	861
Share of non-religious bystanders	0.0	1.0	0.45	0.0	0.48	861
Share of bystanders full-time employed	0.0	1.0	0.38	0.0	0.46	859
Share of bystanders with university education	0.0	1.0	0.36	0.0	0.45	861

[a] The temperature variable measures the absolute temperature (in degrees Celsius) at the specific time and location of the intervention.

[b] All variables are coded at the iteration level. Share women and share of bystanders with earphones were observed by enumerators for all iterations. In 2019, enumerators also estimated bystanders' ages (age brackets) and whether they had an immigrant background or were natives. The variables mean age, share Christian, share non-religious, share full-time employed, and share w/ university draw on putatively unrelated, post-intervention surveys of random samples of bystanders.

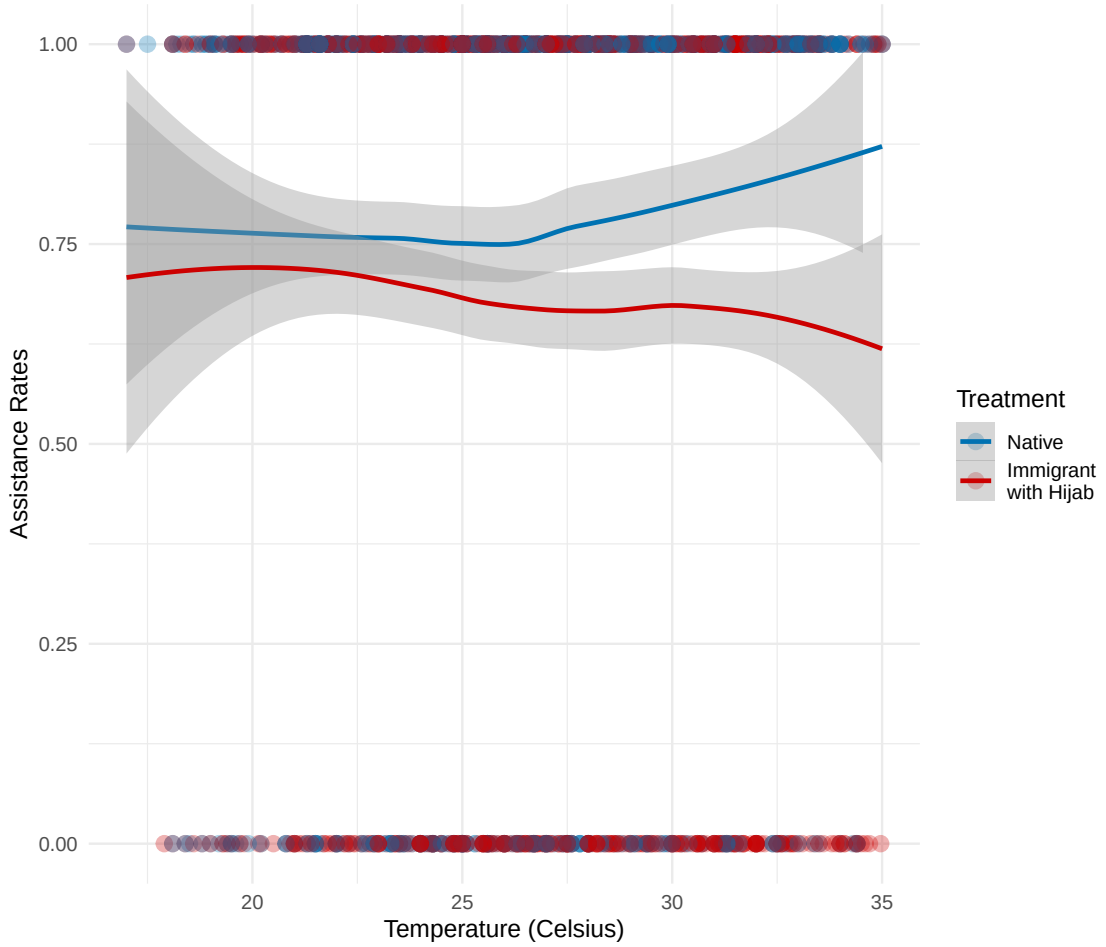

Figure S5: Help rates in response to the two treatments by absolute temperature with LOESS curves. The graphs shows whether any bystander offered assistance to native vs. veiled immigrant confederates by local temperature during the interaction. The shaded areas areas around the fitted LOESS curve show 95% CIs.

Table S2: Help Behavior at Hot Temperatures

	Hijab versus native comparison					
	Outcome: Did any bystanders offer help?					
	(1)	(2)	(3)	(4)	(5)	(6))
Hot Temperature	0.007	0.034	0.049**	0.074***	0.044**	0.065***
	(0.026)	(0.029)	(0.025)	(0.027)	(0.024)	(0.026)
Hijab vs Native	-0.055	-0.049	-0.037	-0.032	-0.049	-0.043
	(0.045)	(0.045)	(0.041)	(0.040)	(0.046)	(0.045)
Hot Temperature x Hijab vs Native	-0.055*	-0.057*	-0.092***	-0.092**	-0.086**	-0.085**
	(0.041)	(0.041)	(0.038)	(0.037)	(0.047)	(0.047)
Constant	0.772***		0.747***		0.753***	
	(0.028)		(0.027)		(0.026)	
Temperature Threshold	24°C	24°C	25°C	25°C	26°C	26°C
Rush Hour FE	No	Yes	No	Yes	No	Yes
Station FE	No	Yes	No	Yes	No	Yes
Observations	1,786	1,786	1,786	1,786	1,786	1,786

[a] Models are estimated with linear regression. Standard errors (clustered at the station level) in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$, one-tailed test.

[b] The temperature variable measures the absolute temperature (in degrees Celsius) at the specific time and location of the intervention.

Table S3: Help Behavior by Temperature: East vs. West

	Hijab versus native comparison		
	Outcome: Did any bystanders offer help?		
	(1)	(2)	(3)
Temperature	0.005*	0.005*	0.006**
	(0.003)	(0.003)	(0.003)
Hijab vs Native	0.138	0.135	0.131
	(0.188)	(0.190)	(0.193)
East	0.167	0.161	0.152
	(0.153)	(0.156)	(0.148)
Temperature x Hijab vs Native	-0.007	-0.007	-0.007
	(0.006)	(0.006)	(0.006)
Temperature x East	-0.004	-0.004	-0.004
	(0.006)	(0.006)	(0.006)
Hijab vs Native	0.015	0.025	0.054
	(0.247)	(0.253)	(0.246)
Temperature x Hijab vs Native x East	-0.004	-0.004	-0.005
	(0.009)	(0.009)	(0.009)
Constant	0.620***		
	(0.101)		
Rush Hour FE	No	Yes	Yes
Number of Bystanders FE	No	No	Yes
Observations	1,786	1,786	1,786

[a] Estimated with linear regression. Standard errors (clustered at the station level) in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$, one-tailed test.

[b] The temperature variable measures the absolute temperature (in degrees Celsius) at the specific time and location of the intervention.