Supporting Information (SI Appendix)
Linguistic Assimilation Does Not Reduce Discrimination Against Immigrants

Donghyun Danny Choi, Mathias Poertner, Nicholas Sambanis

Friday 29th May, 2020

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.

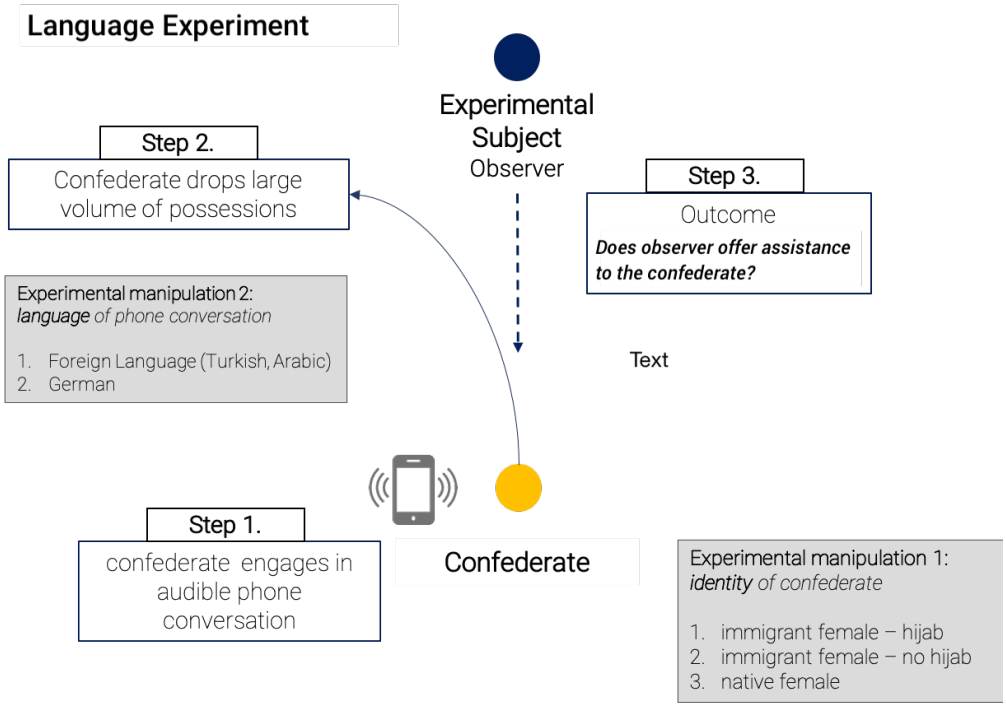


Figure S1: Experimental Setup

Treatment manipulation

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)
- **Dimension 2**: Language in which the confederate conducts the phone call.[1]
    1. Phone call is conducted in a foreign language (Turkish or Arabic)
    2. Phone call is conducted in German

[1]The content of the phone call was an innocuous conversation between friends.

Outcomes

We are interested in measuring the level of assistance offered to the female confederate who drops her possessions (bag of oranges or lemons) 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)
- *bystander_fem*: Total number of female bystanders within the 3 meter radius (count)
- *bystanderHP*: Total number of bystanders with headphones or earphones (count)
- *help*: Whether each bystander offered assistance to the confederate (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.**)

2 Logistics and Procedures

Site selection

The interventions were conducted at 31 train stations across 30 medium to large-sized cities/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 North Rhine-Westphalia (NRW) and Lower Saxony versus the two other states (Brandenburg, and 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 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 largestt 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)
- **Brandenburg**: Potsdam (7), Forst (5), Cottbus (10), Frankfurt-Oder (12), Brandenburg (6)

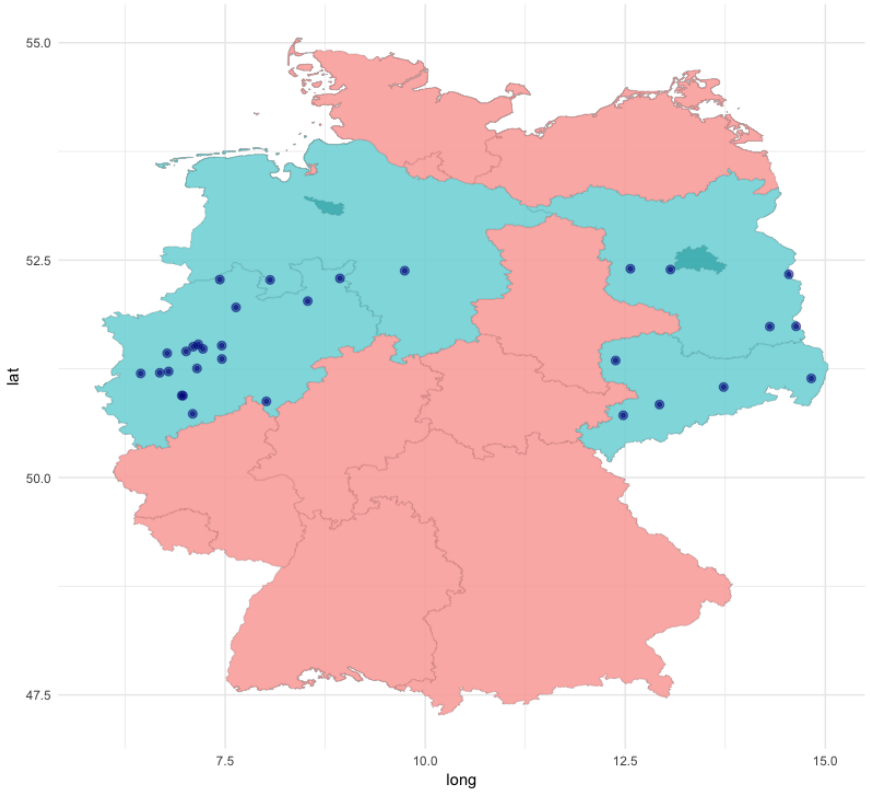


Figure S2: Study sites – 31 train stations across 4 states in North Rhine-Westphalia, Saxony, Lower Saxony, and Brandenburg

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 main 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 University of Pennsylvania's Institutional Review Board (IRB Protocols #829824 and #833206). 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 Bystander Composition and Scene Characteristics

In this subsection, we present descriptive statistics and additional information on the composition of the bystanders and other iteration characteristics. As discussed above, treatment assignment was orthogonal to all bystander characteristics. Therefore, we should not expect these characteristics to affect the results. To further demonstrate empirically that, for example, the number of bystanders does not systemically affect the results, we also report specifications that have number of bystander fixed effects, where the proportion outcome is used in the analysis. The estimates are virtually the same as without the fixed effects. We also include the full set of bystander composition and scene characteristics in our regression-based analyses. As expected, the inclusion of these additional covariates does not change our original findings.

Table S1: Descriptive Statistics on Scene Characteristics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
# of bystanders	1,568	2.969	1.595	1.000	2.000	4.000	10.000
Prop. of bystanders with earphones	1,568	0.055	0.162	0.000	0.000	0.000	1.000
Prop. of female bystanders	1,546	0.554	0.356	0.000	0.333	0.800	3.000
Temperature during iteration	1,474	27.073	4.313	17.000	23.863	30.200	39.700
Rush hour	1,568	0.265	0.441	0	0	1	1

4 Covariate Balance

In this subsection, we present covariate balance statistics for our experimental treatment conditions. While covariate imbalance can arise due to chance, the randomization seems to have successfully obtained balance on each of the 6 pretreatment covariates we collected, both in the full sample as

well as the samples disaggregated by state. Tables S2 and S3 present balance statistics for all statistical tests included in Figures 2 and 3 of the main text. Table S2 presents the balance statistics for the hijab and native comparison. Table S3 presents balance statistics for the hijab-language interaction.

Table S2: Covariate Balance, Figure 2

	Mean Treated	Mean Control	T P-value
Figure 2: Hijab vs No Hijab			
# of Bystanders	2.9461561	2.9823362	0.7068160
Proportion of Bystanders w/ Earphones	0.0601870	0.0477719	0.1906911
Proportion of Female Bystanders	0.5415775	0.5789156	0.1771745
Temperature at Iterations	26.8407277	27.0836752	0.3307750
Rush Hour	0.2552817	0.2786325	0.3705404
Figure 2: No Hijab vs Native			
# of Bystanders	2.9823362	2.9821628	0.9987667
Proportion of Bystanders w/ Earphones	0.0477719	0.0717974	0.0553349
Proportion of Female Bystanders	0.5789156	0.5396544	0.1149995
Temperature at Iterations	27.0836752	27.1085953	0.9360263
Rush Hour	0.2786325	0.2474916	0.3175469
Figure 2: Hijab vs Native			
# of Bystanders	2.9461561	2.9821628	0.7525891
Proportion of Bystanders w/ Earphones	0.0601870	0.0717974	0.3801069
Proportion of Female Bystanders	0.5415775	0.5396544	0.9369361
Temperature at Iterations	26.8407277	27.1085953	0.3909463
Rush Hour	0.2552817	0.2474916	0.8015982

Table S3: Covariate Balance, Figure 3

	Mean Treated	Mean Control	T P-value
Figure 3: Hijab Foreign Language vs German			
# of Bystanders	3.1370656	2.7861381	0.0123732
Proportion of Bystanders w/ Earphones	0.0593355	0.0609007	0.9149576
Proportion of Female Bystanders	0.5399918	0.5429067	0.9190942
Temperature at Iterations	26.6527671	26.9982740	0.3348980
Rush Hour	0.2702703	0.2427184	0.4555253
Figure 3: No Hijab Foreign Language vs German			
# of Bystanders	2.9408213	3.0194175	0.5527557
Proportion of Bystanders w/ Earphones	0.0386188	0.0559475	0.1467414
Proportion of Female Bystanders	0.5834871	0.5748323	0.7791564
Temperature at Iterations	26.8794686	27.2660734	0.2695371
Rush Hour	0.2789855	0.2783172	0.9856705
Figure 3: Hijab German vs Native			
# of Bystanders	2.7861381	2.9821628	0.1288787
Proportion of Bystanders w/ Earphones	0.0609007	0.0717974	0.4719210
Proportion of Female Bystanders	0.5429067	0.5396544	0.9091103
Temperature at Iterations	26.9982740	27.1085953	0.7532378
Rush Hour	0.2427184	0.2474916	0.8914249

5 Additional Analyses

Next, we present results disaggregated by experiment. As mentioned in the main text, the experiments used common treatment arms, so we pool across the two experiments in the main analysis. The analysis disaggregated by experiment reveals similar results. We first present results from the summer 2018 experiment and then from the 2019 experiment. In neither experiment do we observe statistically significant differences in discrimination generated by the perception of linguistic assimilation. The comparison of help rates in columns 1 vs 2 and 3 vs 4 is not statistically significant in either experiment.

Effects disaggregated by experiment

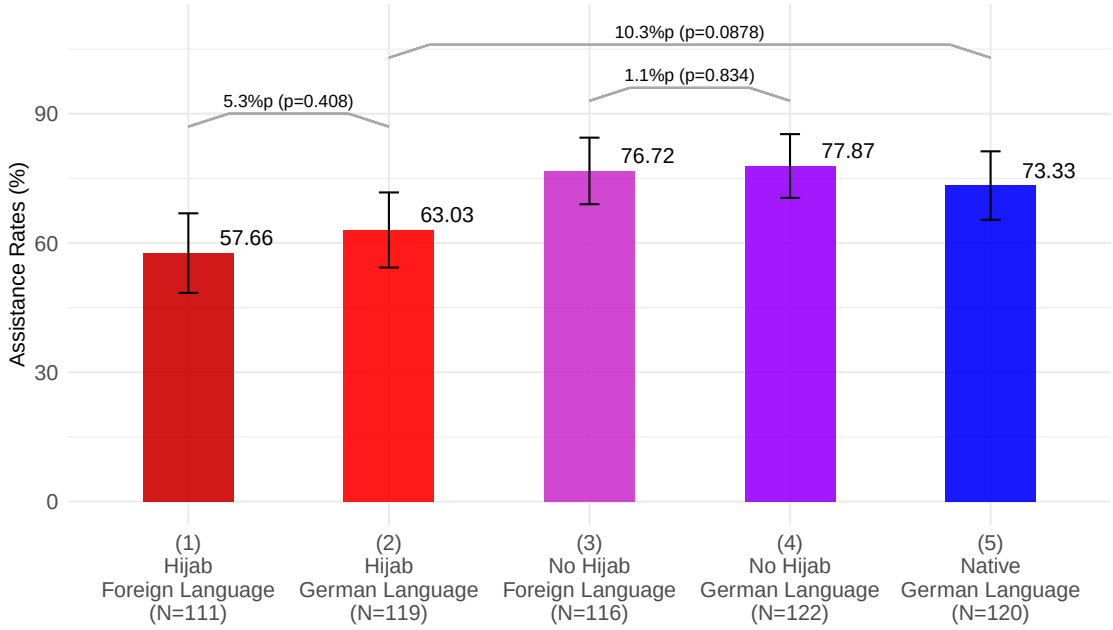


Figure S3: Language Effects: Experiment 1 (Summer 2018)

The bars reflect the mean rate of assistance for each of the treatment conditions, with 95% confidence intervals. The lines that connect the bars are from a two-tailed difference-in-means tests of the conditions, with associated p-values. The figure draws from data from experiment 1 (Summer 2018) only.

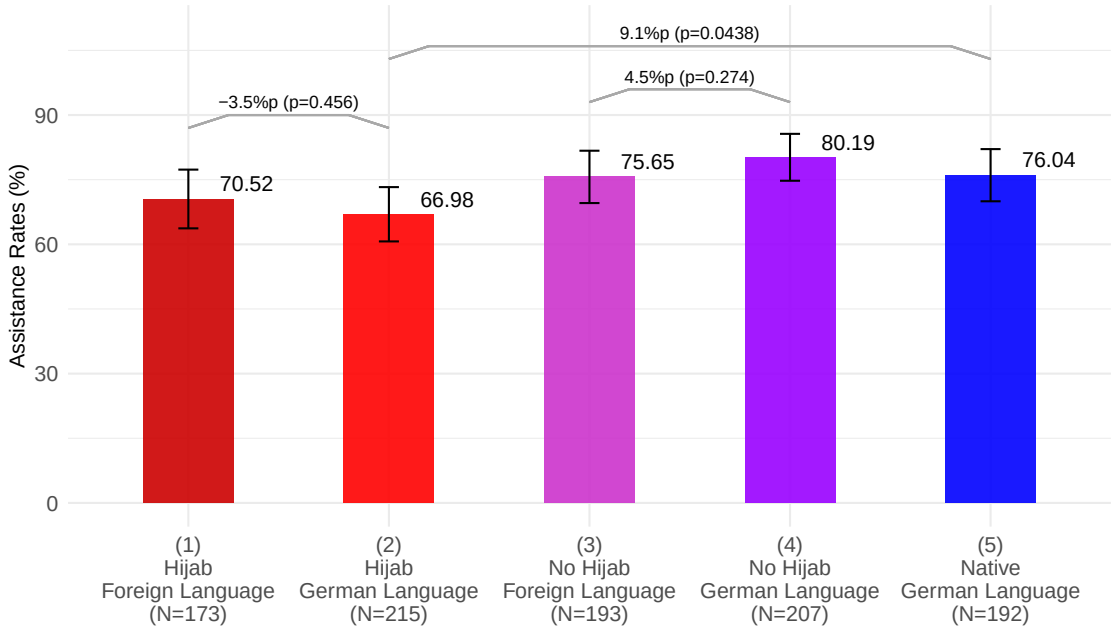


Figure S4: Language Effects: Experiment 2 (Summer 2019)

The bars reflect the mean rate of assistance for each of the treatment conditions, with 95% confidence intervals. The lines that connect the bars are from a two-tailed difference-in-means tests of the conditions, with associated p-values. The figure draws from data from experiment 2 (Summer 2019) only.

Effects disaggregated by region: Former West vs East Germany

Next, we present results from the analysis disaggregated by region. Immigrant population density is lower in the East than the West and anti-immigrant sentiment is higher in the East. We test if perceived linguistic assimilation has different size effects in one of these regions, but find no such evidence.

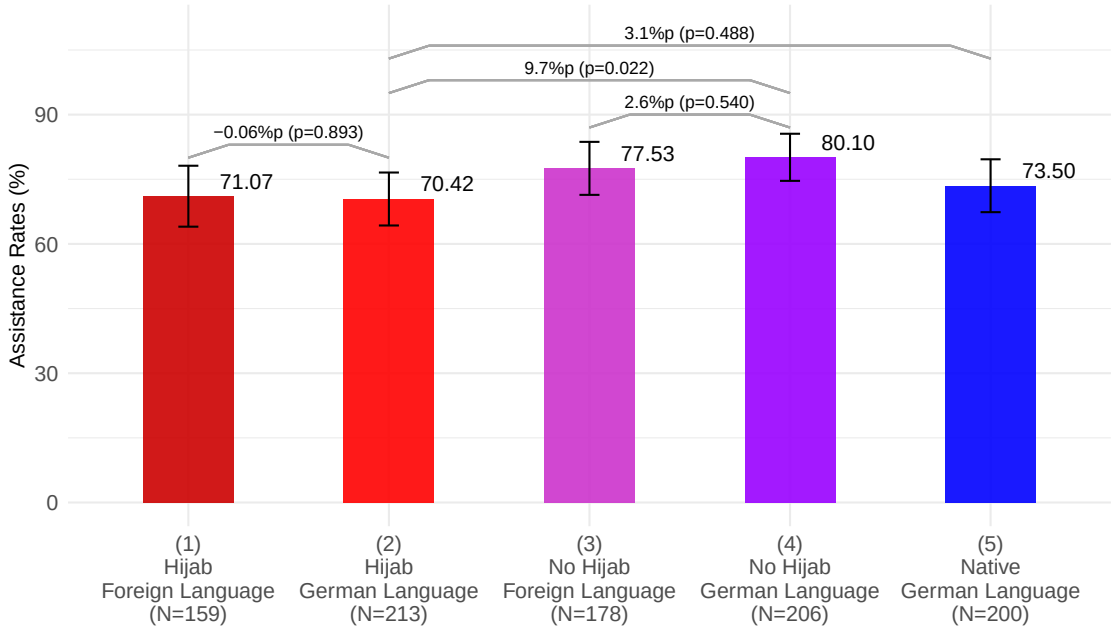


Figure S5: Language Effects: Former West Germany

The bars reflect the mean rate of assistance for each of the treatment conditions, with 95% confidence intervals. The lines that connect the bars are from a two-tailed difference-in-means tests of the conditions, with associated p-values. The figure pools data across experiments 1 (Summer 2018) and 2 (Summer 2019), but subsets to data from German states that belonged to former West Germany (Northrhein-Westfalia and Lower Saxony).

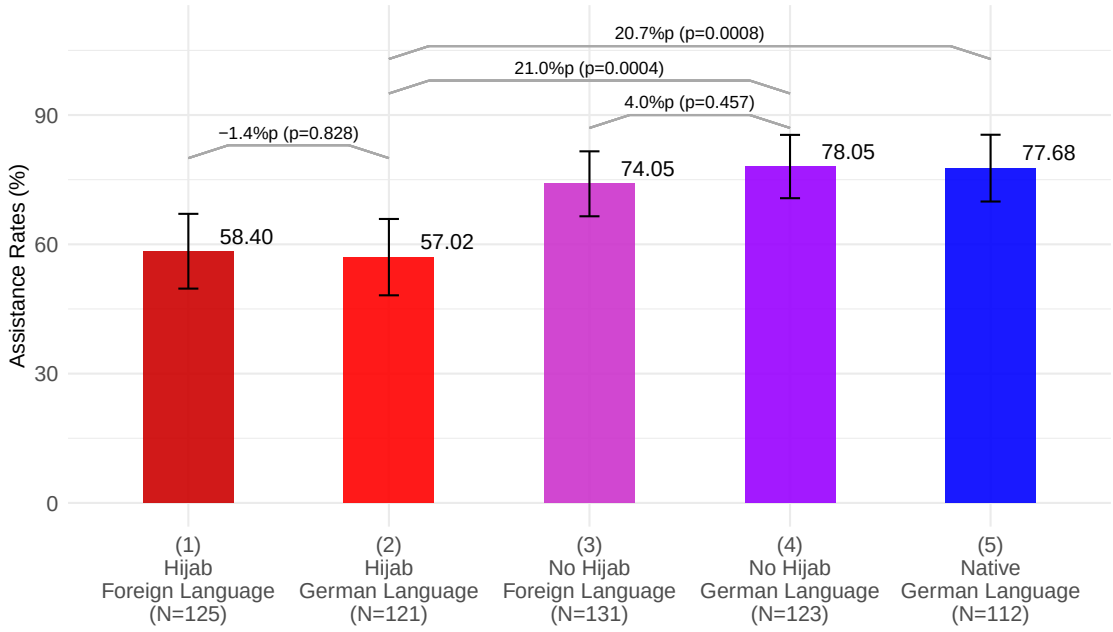


Figure S6: Language Effects: Former East Germany

The bars reflect the mean rate of assistance for each of the treatment conditions, with 95% confidence intervals. The lines that connect the bars are from a two-tailed difference-in-means tests of the conditions, with associated p-values. The figure pools data across experiments 1 (Summer 2018) and 2 (Summer 2019), but subsets to data from German states that belonged to former East Germany (Brandenburg and Saxony).

6 Equivalence Tests

The analysis in the main paper has demonstrated that linguistic assimilation fails to reduce discrimination against Muslim immigrants; across a range of approaches, our analyses have shown that we are consistently unable to reject the null hypothesis that the true effect of linguistic assimilation on discrimination is zero at conventional levels. Skeptics might be concerned that our inability to reject the null hypothesis does not mean that the null hypothesis is statistically supported; it might potentially be the case that our failure to reject the null hypothesis is the result of a lack of statistical power to detect a true effect that exists and is positive, rather than the true effect being zero.

We note that our experiments are sufficiently well-powered to detect discrimination along a different dimension (religion) and, according to prior literature, linguistic differences should have had a *larger* effect than any other dimension of cultural difference. Moreover, the coefficient estimates for bias due to language differences are so small as to suggest that there is no substantively important effect. To further reinforce this argument, we provide additional evidence below that null effects on linguistic assimilation should be interpreted as an *absence* of a *substantively meaningful* effect using a series of equivalence tests. (Berger, Hsu et al., 1996; Seaman and Serlin, 1998; Wellek, 2010). We set equivalence bounds based on the size of the discrimination effect due to religious difference and present results using different bounds.

While demonstrating that the true effect of an experimental treatment is precisely zero is impossible, scholars in the frequentist paradigm of hypothesis testing have devised statistical tests that allow us to reject that the treatment effects are large enough to be of substantive import. In the so-called "equivalence testing" paradigm, scholars are able to statistically reject that the treatment effects are more extreme (larger in magnitude) than a *predetermined* upper and lower threshold at which the magnitude of the effect are deemed "large", and thus consider the true effect to be close enough to zero for practical purposes.

To find evidence *against* large treatment effects of linguistic assimilation, we adopt this equivalence testing approach, and the Two One-Sided Tests (hereafter TOST) in particular. TOSTs has gained favor amongst scholars in the biomedical and psychological sciences for being a simple and intuitive approach to demonstrating equivalence. The logic of the TOSTs are as follows; first, the researcher must set an upper (B_U) and lower (B_L) equivalence bound, based on the smallest effect size of interest. Then, two null hypotheses—namely that 1) $B_T \leqslant -B_L$ and 2) $B_T \geqslant B_U$. By showing that these statistical tests can be rejected, we are able to claim that the observed effect falls within the equivalence bounds $B_L \leqslant B_T \leqslant B_U$, and thus no longer meaningfully different from a negligible effect.

Since our outcomes are measured dichotomously, we apply the Two one-sided test of *proportions* framework to our analyses. In order to do so, we need to specify equivalence bounds, or more intuitively, the threshold at which we would consider the effect to be no longer of substantive import. Setting these equivalence bounds is an inherently arbitrary practice, and the literature is divided as to what the most appropriate approach is. We make a deliberate choice to set equivalence bounds on the linguistic assimilation effects based on the magnitude of the discrimination observed in assistance rates between immigrant confederates and native confederates (12.0%p) that we documented in a peer-reviewed published study. For transparency's sake, we present results for equivalence tests that set equivalence bounds at 33% (1/3, 4%p), 50% (1/2, 6.0%p), 66% (2/3, 8%p) of the total discrimination effect. Given the extensive literature that predicts the importance of linguistic differences in driving discrimination, we find the parameters we use to be reasonable. It is also interesting to note that

Table S4: Equivalence Tests for Linguistic Assimilation Effects

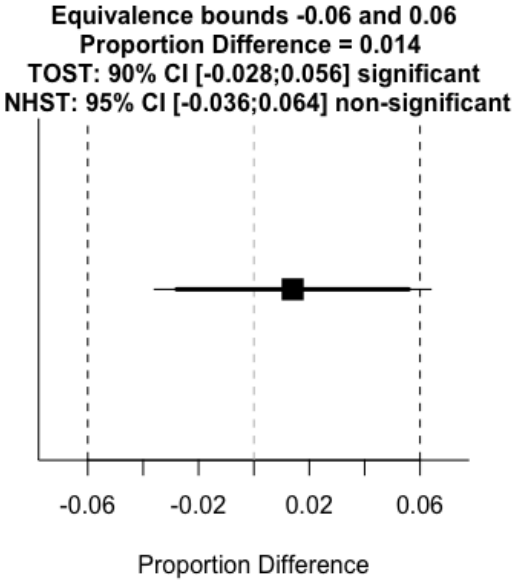
	Equivalence Bound	TOST Conf. Interval	Fisher's Exact Z Test
	Test 1: Immigrant Foreign vs German Language (ATE: 1.4%p)		
Test 1 small (33%)	(-0.04, 0.04)	(-0.028, 0.056)	p = 0.154
Test 1 intermediate (50%)	(-0.06, 0.06)		p = 0.035
Test 1 large (66%)	(-0.08, 0.08)		p = 0.005
	Test 2: Hijab Foreign vs German Language (ATE: 0.07%p)		
Test 2 small (33%)	(-0.04, 0.04)	(-0.062, 0.064)	p = 0.152
Test 2 intermediate (50%)	(-0.06, 0.06)		p = 0.061
Test 2 large (66%)	(-0.08, 0.08)		p = 0.019

the 6%p bounds are roughly equivalent to the recommendation made by Simonsohn (2015) to set the equivalence bounds to the effect size that would have given a study 33% power.

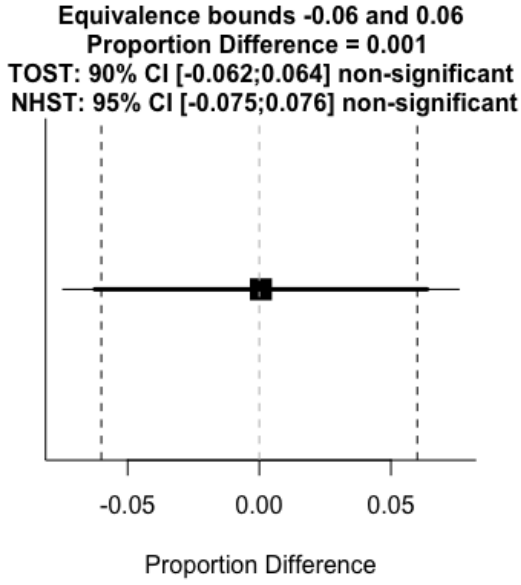
We report the results of the TOSTs in Table S4 and Figure S7. Whereas we are unable to claim equivalence when we set the equivalence bounds to 1/3 of the total discrimination effect (4% points, Fisher's exact Z test p value = 0.154), we are able to claim, based on the equivalence test and the original null-hypothesis tests, that the observed effect for linguistic assimilation is statistically not different from zero, and statistically equivalent to zero; the confidence interval for the Two One-sided Tests falls within the equivalence bounds for intermediate and large bounds (6%, 8%points) and the p-values for the Fisher's Exact Z test fall below conventional levels (p=0.035, p=0.004 respectively). When we conduct equivalence tests for linguistic assimilation effects for hijab-wearing immigrant confederates, we observe similar patterns (p=0.061, p=0.019 respectively).

We interpret these findings as statistical evidence against substantively large treatment effects of linguistic assimilation. Combined with the analyses presented in the main text of the paper, we are confident that linguistic assimilation is likely to have negligible influence on reducing discrimination against Muslim minority immigrants.

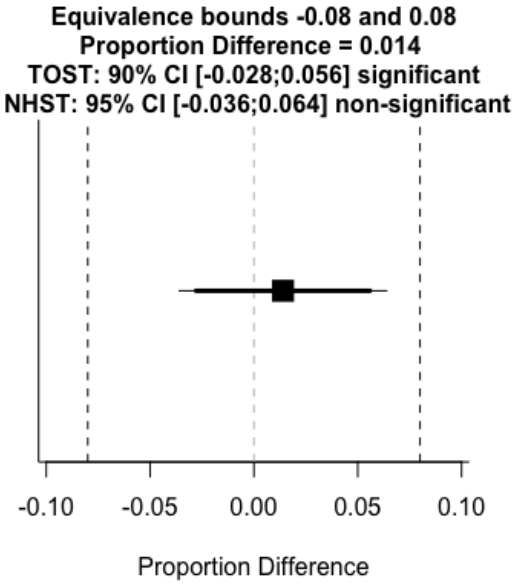
Figure S7: Equivalence Testing: Two One Sided Test of Proportions



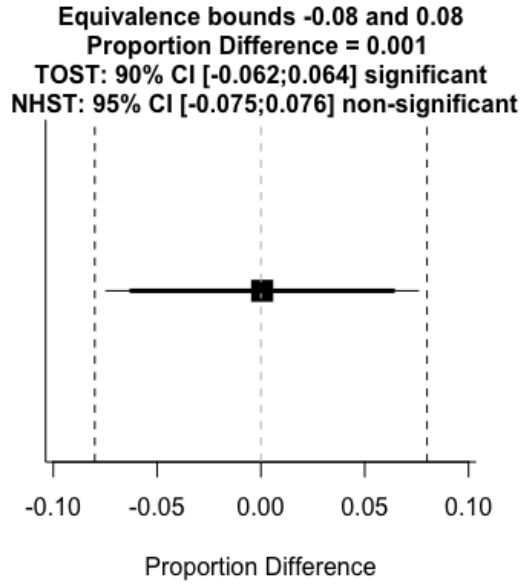
(a) Test 1, Equivalence Bounds (-0.06, 0.06)



(b) Test 2, Equivalence Bounds (-0.06, 0.06)



(c) Test 1, Equivalence Bounds (-0.08, 0.08)



(d) Test 2, Equivalence Bounds (-0.08, 0.08)

Note: The filled squares represent the point estimate for our observed effects, with the thick and thin lines representing 90% confidence interval for the TOST (equivalence test), and the 95% confidence interval for the null hypothesis significance tests respectively. The horizontal dotted lines represent the equivalence bounds that we set for the TOSTs; $\pm$ 6% points for subfigures (a) and (b) and $\pm$ 8% points for subfigures (c) and (d). TOSTs were implemented using the `TOSTER` package in `R`.

7 Effects by Foreign Language Used

In this section, we examine whether there exist any heterogeneity in the effects of linguistinc assimilation for hijab-wearing immigrant confederates. We specifically examine whether the type of foreign language used by the immigrant confederate drives any heterogeneity. In order to do so, we compare the language effects by whether the confederate spoke Arabic or Turkish during the phone call. Columns (1) and (2) of Table S5 subsets the sample to find that there are no statistically significant linguistic assimilation effects (ATE = 0.8%p and -2.9%p respectively). In column (3), we report the interaction term between our Foreign vs German language treatment with an indicator variable for Turkish language iterations. While the Turkish language iterations result in marginally higher assistance rates than Arabic language iterations, this difference falls far short of statistical significance.

Table S5: Effects by Foreign Language Used

	Dependent variable:		
	Any help?		
	(1)	(2)	(3)
Foreign vs German (Hijab)	0.008	−0.029	−0.029
	(0.043)	(0.084)	(0.082)
Turkish			0.034
			(0.064)
Foreign vs German × Turkish			0.037
			(0.093)
Constant	0.663***	0.629***	0.629***
	(0.029)	(0.059)	(0.057)
Foreign Language	Turkish	Arabic	Merged
Observations	483	135	618
R^2	0.0001	0.001	0.002
Note:			*p<0.1; **p<0.05; ***p<0.01

8 Manipulation Checks on Perception of Confederate Ethnicity

In this section, we demonstrate that German native populations accurately recognize that our minority confederates to be of immigrant minority background (in the control condition when they are *NOT* wearing a hijab). In order to do so, we conducted a follow-up survey on Clickworker.com, an online crowdsourcing work platform similar to Amazon's M-Turk to recruit adult German respondents to evaluate our confederate's photos and report their perceived country of origin. We conducted this survey on a sample of 208 German adults above 19 years of age. Each evaluation question presented a photo of our confederate, and then asked "in your best guess, where do you think this person is from?" Respondents were then asked to choose from "German" versus four other countries (Turkey, Egypt, Iraq, and Syria), which were the real countries of origin for our immigrant confederates. All respondents evaluated a total of 15 confederate photographs, and roughly 1/2 of the total German native confederates that participated in the intervention of the experiment. This yields a total of 3,120 evaluations across all photos. A screen capture of a typical evaluation task for the manipulation check are presented in Figure S8.

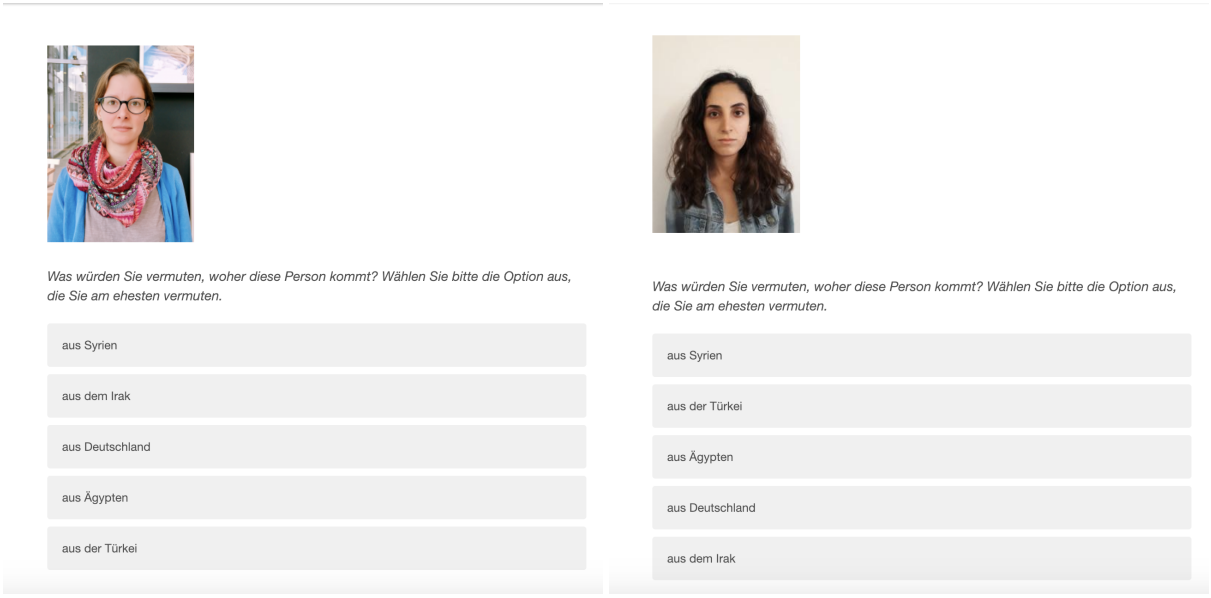


Figure S8: Screen Capture of Manipulation Check Task

It is clear that respondents are able to draw stark distinctions in the country of origin of our German native confederates versus immigrant confederates. On average, respondents correctly identify German native confederates as Germans between 82–83% of the time. In stark contrast, only 15–16% of respondents mistakenly categorize our immigrant minority confederates' country of origin as Germany. The difference is consistently in excess of 65% points, and is statistically distinguishable at $p<0.001$. These manipulation checks provide strong evidence that our immigrant confederates were sufficiently different in terms of their ethnic attributes (phenotype, skin tone) to German native confederates, and bystanders in our main experiment are highly likely to have perceived our immigrant control confederates as immigrants or Germans with an immigrant background. As with every survey, it is possible to consider different ways of presenting the survey questions. For example, a

longer list of countries could have been provided to respondents to choose from; other countries (beyond Germany) with majority Christian population could have been included; or responses could have been left open-ended. Nonetheless, the evidence in this survey is so stark as to suggest that these slight modifications would not impact our conclusions from the manipulation checks.

Table S6: Proportion of respondents identifying confederate as a German native

Native Confederates	Immigrant Confederates	Difference	P-Value
82.99%	16.74%	66.246%p	< 0.001

References

Berger, Roger L, Jason C Hsu et al. 1996. "Bioequivalence trials, intersection-union tests and equivalence confidence sets." *Statistical Science* 11(4):283–319.

Seaman, Michael A and Ronald C Serlin. 1998. "Equivalence confidence intervals for two-group comparisons of means." *Psychological Methods* 3(4):403.

Wellek, Stefan. 2010. *Testing statistical hypotheses of equivalence and noninferiority*. Chapman and Hall/CRC.