sociological science

Supplement to:

Auspurg, Katrin, Renate Lorenz, and Andreas Schneck. 2023. “Does Unprecedented Mass Immigration Fuel Ethnic Discrimination? A Two-Wave Field Experiment in the German Housing Market.” Sociological Science 10: 640-666.

Online Supplement for

Does Unprecedented Mass Immigration Fuel Ethnic Discrimination? A Two-Wave Field Experiment in the German Housing Market

Katrin Auspurg[a,b], Renate Lorenz[a] & Andreas Schneck[a,b]

[a] LMU Munich, Department of Sociology; Konradstr. 6, 80801 Munich, Germany
[b] Goethe-University Frankfurt a. M., Faculty of Social Sciences, 60323 Frankfurt a. M., Germany
Contact: katrin.auspurg@lmu.de, renate.lorenz@soziologie.uni-muenchen.de, andreas.schneck@lmu.de

Content

S1. Materials and Methods

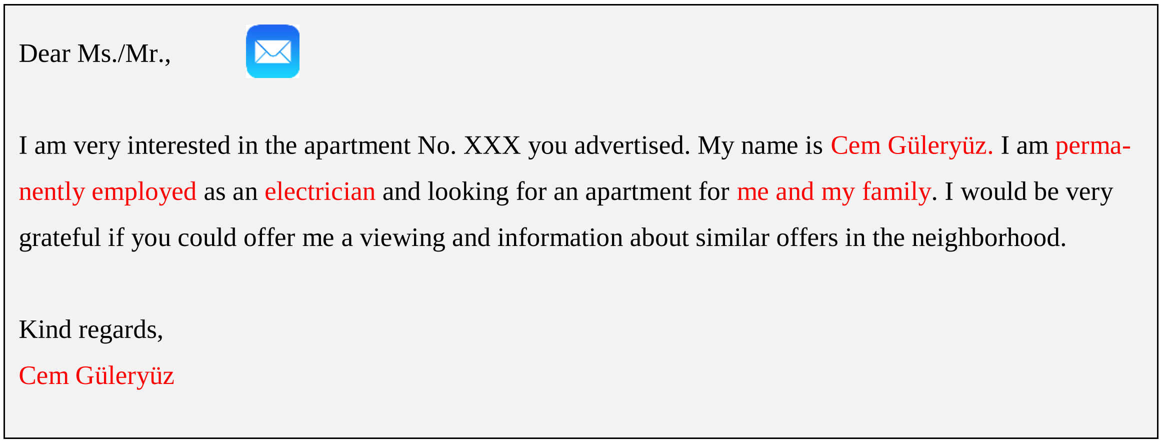
S1.1. Experimental Data

<u>*Field Experiment Design (E-mail Correspondence Test)*</u>

The main treatment variable of the e-mail correspondence test was applicant ethnicity (Turkish or German). We used a paired testing design: Each housing supplier in the sample received one request for a viewing appointment from a male Turkish applicant and one request from a male German applicant. We used 30 common first and last names to signal each ethnicity. Only male first names were used, as we were only interested in ethnic discrimination, not gender discrimination. One name was randomly selected from the pool of 30 names from each ethnic background. The names were included in the application as well as in corresponding e-mail addresses set up with common providers (such as cem.gülerüz@web.de; carsten.schweiger@gmx.de). Using many different names per ethnicity helps to avoid confounders, such as age or social background (Gaddis 2017). The absence of such confounding factors was indicated by the high effect homogeneity observed among the various names signaling Turkish (vs. German) ethnicity (see Section S3.2, Figure S7). All applications were written in correct standard German. To avoid order effects and to achieve a balanced design, the e-mails were sent in alternating order within a time interval of approximately one hour. In addition to ethnic background, we varied some other applicant characteristics as experimental factors. (In the present project, these additional applicant characteristics serve only as control variables to standardize the applicant profiles.)

All applicant characteristics were fully crossed based on a *D*-efficient experimental design, the gold standard for optimal orthogonal and balanced experimental designs (Auspurg and Hinz 2015). *D*-efficient designs minimize the correlations (possible confounding) between experimental factors and maximize their variance to obtain maximum statistical power to identify and separate the effects of all experimental factors. While ethnicity was always varied between the two applications to the same housing unit (i.e., one application was always sent by a Turkish, one by a German applicant), the levels of other experimental factors could either be the same or differ between the two applicants applying to the same housing unit. Such a design prevents confounding of the experimental stimuli with the composition of the applicant pool applying to the same unit (Phillips 2019). At the same time, it conceals the nature of the experiment from housing suppliers while taking advantage of the high internal validity of a paired testing design.

As is common in the German rental housing market, we consistently used brief e-mail queries that included only the most important information about the applicant's background.[1] To minimize the risk of the experiment being recognized by suppliers, a slightly different wording was used for the two applications sent to the same housing unit. These slightly different text versions (e.g., using different salutations or orders of text phrases) were randomly assigned to the applications. Figure S1 shows an example of an e-mail request with the experimentally varied applicant characteristics highlighted in red. A descriptive overview of all experimental factors and levels, along with balance checks, is provided in Section 3.2.



Dear Ms./Mr.,

I am very interested in the apartment No. XXX you advertised. My name is Cem Güleryüz. I am permanently employed as an electrician and looking for an apartment for me and my family. I would be very grateful if you could offer me a viewing and information about similar offers in the neighborhood.

Kind regards,
Cem Güleryüz

Figure S1. Example of an e-mail request for an apartment viewing, translated from German. The experimentally varied factors are highlighted in red.

[1] We are very confident that we did not miss any key features of standard applications in the German housing market: The housing platform also asked for this core information in a small pilot with tabular application forms. This pilot was implemented by the platform shortly before the 1st wave of our field experiment (while during our field period, we could just apply with unstructured e-mails). To ensure that we used standard applications, we also advertised a (hypothetical) apartment ourselves when preparing the materials for our field experiment. Thereby, we were able to collect many sample e-mails from housing applicants on which to base our text versions.

The resulting pairs of e-mails were randomly assigned to the sampled housing units. To minimize errors, we used an automatic web-scraping procedure to send the e-mails, with a time gap of approximately one hour between the two applications (with some small random time variation to conceal the nature of the experiment).

If a housing supplier responded to an application with an offer to visit the housing unit, we politely declined within a short period (under the pretense of already having found a rental) to minimize the burden on the suppliers. This was in agreement with the Ethics Committee that approved our field experiments on the German housing market.

Sample of Tested Housing Units

All housing units tested in this field experiment were sampled from a major online housing platform listing private and corporate advertisements. We exclusively sampled rentals, because discrimination is likely to be more prevalent in the rental housing market than in the real estate market.[2] Another reason for restricting the sample to rental units was that in Germany, especially among migrants, rental housing is far more common than homeownership (German Federal Statistical Office - Destatis 2021).

The two sampling periods took place in spring and winter 2015 (1st wave: May 4th – May 8th, 2nd wave: November 30th – December 4th). During both waves, a random sample of 2,500 advertised rental housing units with 2 to 4 rooms was drawn (500 units per day), resulting in a total sample size of 5,000 units. For ethical reasons and to follow standard procedures for field experiments in the housing market, we sampled on the level of suppliers and not housing units so that we tested each supplier only once. After concluding the sampling procedures, a few housing units (N = 188) were excluded from the analysis sample as they were no longer available on the housing platform at the time the e-mail of the 2nd applicant was planned to be sent. In these cases, a paired test was not feasible. A few more units (N = 13) were excluded, as no information on their regional location was available, making it impossible to measure moderator variables (i.e., share of foreigners living

[2] In the rental market, providers must be confident that tenants will care for the rental and make reliable rent payments. Economic theories suggest that this setting provides stronger incentives for ethnic discrimination than the real estate market, where there is typically only a one-time financial transaction.

in the region) or control variables (e.g., the federal state in which the housing unit was located).

The analysis sample included 4,799 tested rental units: 2,389 in the 1st and 2,410 in the 2nd wave. In both waves, the apartments were distributed throughout Germany. Overall, the field experiment was run in all federal states and in 388 of the 401 counties in each of the two waves.

Main Treatment Variable: Refugee Immigration

Our main treatment variable for identifying the impact of refugee immigration on discrimination is the timing of our experiment: shortly before the beginning of the refugee crisis (1st wave) or at the peak of the crisis (2nd wave). In robustness checks, we use a metric measure of the refugee immigration to different counties as an alternative treatment: The magnitude of immigration into different counties ranged from 0.3 to 1.2 refugees per 100 inhabitants (counties with large reception centers excluded). As a further treatment variable, we were able to calculate the walking distance from the tested housing units to (newly established) refugee shelters for several federal states (see Section S3.1 on robustness checks).

Moderator and Control Variables

In terms of treatment effect heterogeneity, we examine the moderation of the effect of refugee immigration by the size of the immigrant population that already lived in a county prior to the refugee crisis. We measure the immigrant population size with the share of foreigners per county, i.e., the share of persons without German citizenship in the year before the refugee crisis set in (2014). Note that this share varied strongly among the observed counties (from 1.0% to 32.3%), as non-refugee migrants are allowed to move freely across Germany. The Federal Statistical Office provides the numbers of foreigners on an annual basis. The number of foreigners is strongly correlated with the number of residents with an immigrant background (i.e., residents, who themselves or at least one of their parents were born without German citizenship).[3] We also use official statistics from the German Federal Statistical Office at the county level for other regional context characteristics, such as vacancy rates, which are used as controls in robustness analyses.[4]

S1.2. Analytical Strategy

The aim of this paper is to identify the impact of refugee immigration on discrimination. Therefore, we must first measure discrimination; and second, we must determine whether the extent (S1.2.2) of discrimination changed as a result of immigration.

Discrimination Rates

To measure discrimination, we use differences in the suppliers' reactions. The most important difference in their reaction is whether an applicant receives a response, as receiving a response is a precondition for continuing the application process.[5] Differences between paired responses are analyzed to investigate whether Turks were treated differently than Germans. This strategy allows us to draw on the high internal validity of paired testing designs: By contrasting applications to the same rental unit, all housing unit and supplier characteristics are naturally held constant (Vuolo, Uggen and Lageson 2018). In total, there are three different outcomes *j* of interest in a paired design:

$j = 0$: both receive a response, or both do not receive a response

$j = 1$: only the German applicant receives a response

$j = 2$: only the Turkish applicant receives a response

[3] According to the 2011 Census, 7.7% of the population in Germany were foreigners and 19.2% had a migration background, with both shares being correlated at the county level with $r = 0.92$.

[4] For such control variables, we can also use characteristics collected on the housing platform itself. During the web-scraping process, we collected information on all advertisements posted on the platform (not only those tested). This information allowed us to measure, for example, the average listing duration in the different counties (an alternative variable to the vacancy rate to measure the supply-demand situation in local housing markets) and also to collect information on the size of the suppliers (number of listings per supplier). These variables can be used to control for possible shifts in the composition of the sample or the housing market situation between waves. Extensive analyses showed that they neither affect the measured gross nor net discrimination in a substantial way (see the methods analyses reported in Auspurg, Schneck and Thiel 2020).

[5] All requests that did not receive a response within 14 days were coded as non-response.

While the first instance ($j = 0$) represents equal treatment, the two latter instances represent unequal treatment: $j = 1$ (2) indicates an unfavorable treatment of the Turkish (German) applicant.

In standard literature, the proportions or probabilities of unfavorable treatment are defined as the gross discrimination rates of the Turkish and German applicants, respectively:

Gross discrim. rate of Turks: $P_{j1} = P(j = 1)$ (probability that only the German applicant receives a response)

Gross discrim. rate of Germans: $P_{j2} = P(j = 2)$ (probability that only the Turkish applicant receives a response)

These discrimination rates allow us also to identify cases in which the majority applicant (here: German) is disadvantaged compared to the minority applicant (here: Turkish).[6] To quantify the extent to which Turks are systematically more often disadvantaged, i.e. "discriminated" compared to Germans, the net discrimination rate is standard:

Net discrim. rate of Turks: $P_{j1} - P_{j2}$ (difference in the gross discrimination rates)

To test whether the difference between gross discrimination rates is statistically significant, a nonparametric McNemar's (MN) test can be used (Vuolo, Uggen and Lageson 2016, Vuolo, Uggen and Lageson 2018). This test statistic contrasts the gross discrimination rates of Turkish and German applicants: $MN = (P_{j1}- P_{j2})^2/(P_{j1} + P_{j2})$. With a sufficient number of cases, this test follows a χ^2(df=1)-distribution.

Identification of the Effect of Refugee Immigration on Discrimination Rates

We are interested in the following question: Was the level of discrimination systematically affected by the timing of our experiment (i.e., before/at the peak of the refugee crisis)? To answer this question, we use multinomial logistic regressions with the three individual results j of the experiment as the outcome variable. The main predictor is the timing of the experiment (2^{nd} versus 1^{st} wave). (In robustness checks, we use the magnitude of refugee immigration on the county level as an alternative treatment.)

[6] Discrimination of the majority applicant may result, for example, from suppliers having an immigrant background combined with an in-group preference. Another possible explanation is discrimination based on assumed customers' preferences: In neighborhoods with many foreigners, suppliers might assume that Turkish migrants more likely rent an appartment.

Equation (1) shows the regression model (for details on multinomial regression models see Greene 2012: p. 763). Logit specifies the log-transformed odds of the two discrimination outcomes (j = 1 or j = 2) against the reference category of equal treatment (j = 0). i is an index for the tested housing unit (i = 1, …, $N_{\text{housing units}}$). I represents refugee immigration, measured either with the timing of the experiment (1st or 2nd wave) or in robustness checks with our alternative metric measurement of immigration. C are control variables that might affect the level of discrimination (e.g., percentage of foreigners in the county; city vs. rural region).

$$Logit(Y_i = j) = \beta_{0j} + \beta_{Ij}I_i + \beta_{Cj}C_i, \quad j = 0, 1, 2; \quad i = 1, \ldots, N_{\text{housing units}} \tag{1}$$

Positive (negative) regression coefficients mean that the odds of the outcome j is increased (decreased) compared to the reference category. For our research goal, the coefficient β_{Ij} in equation (1) is of most interest: A significant positive coefficient β_{I1} (β_{I2}) would suggest that the level of gross discrimination of Turkish (German) applicants increased across the two waves of our experiment (respectively, due to larger refugee immigration).[7] To simplify interpretation, effects are converted to average marginal effects (using the *margins* command in Stata), which indicate effects on the predicted probability of the respective outcome (gross discrimination rate) in percentage points, averaged across the entire estimation sample.[8]

If both ethnicities showed similar changes in gross discrimination between waves, this would indicate a general trend in the housing market. For instance, a changing relation between supply and demand may affect the response probability similarly for both ethnicities (equal treatment). In contrast, for our purpose, it is particularly interesting whether the gross discrimination rates changed differently for Turkish and German applicants, as this

[7] Multinomial regressions estimate different regression coefficients for each outcome. Here, this allows us to see whether the explanatory factors for the discrimination against the Turkish versus German applicants differ. This is, for example, interesting to see whether housing providers try to steer migrants toward other migrants. In this case, we would observe lower (higher) gross discrimination of Turks (Germans) in regions with a higher share of foreigners.

[8] For example, a reported average marginal effect of 0.05 for the outcome j = 1 (gross discrimination rate of Turks) would mean that the respective variable increased the average predicted probability that only the German applicant received a response (= gross discrimination of Turks) by 5.0 percentage points (implying a decline of 5.0 percentage points in the summarized probability of the two other outcomes, only the Turkish applicant got a response, or equal treatment; we multiply effects by 100 to report effects in percentage points).

would imply a change in the net discrimination rate of Turks. To identify such disproportional change in gross discrimination rates, we test the following null hypothesis (with standard χ^2-tests for the equivalence of marginal effects using the `test` command in Stata):

$$\beta_{I1} = \beta_{I2} \leftrightarrow \beta_{I1} - \beta_{I2} = 0 \qquad (2)$$

A positive difference would mean that over the course of the refugee crisis, the gross discrimination of Turks increased more strongly (or decreased less) compared to the gross discrimination rate of Germans, meaning that the gap (the net discrimination rate of Turks) increased.

In the supplemental analyses, we extend our parsimonious model presented in the main text that does not include any control variables by controlling on the federal state fixed effects, regional (city vs. countryside), advertisement characteristics (private supplier (yes/no), number of rooms, rent per sqm) as well as other variables on the county level (share of foreigners (2014) unemployment rate (2014, 2015), population density (2014, 2015), GDP of employed (2014, 2015), the vacancy rate (2011, census data) as well as the voter share for the green party in the previous federal election (2013)).

It is especially important to control on federal states as they differ in population size and tax revenues, the two parameters used to determine the size of refugee immigration by the so-called “Königsteiner Schlüssel”. By controlling for federal state and this large bundle of control variables, possible imbalances in the composition of housing suppliers or characteristics of regional units (for federal states: even in stable unobserved characteristics) that might confound our treatment effect are leveled out (achieving “conditional ignorability,” see the main text for details on the identification strategy).

We also use nonparametric analyses such as local polynomial smoothing to capture non-linear and threshold effects (e.g., discrimination rates might change more strongly once refugee immigration exceeds a particular threshold).

Identification of Treatment Effect Heterogeneity

We are also interested in possible treatment effect heterogeneity across regions accustomed to varying levels of immigration before the refugee crisis started. This treatment effect heterogeneity is identified by including interaction terms between refugee immigration I

and the share of foreigners in a county as a regional context characteristic F ($I·F$) (see equation 3). A significant coefficient β_{IFj} would indicate a significant change in how regional characteristics moderate the gross discrimination rate. Depending on the levels of these variables, different changes in gross discrimination rates are estimated:

$$Logit(Y_i = j) = \beta_{0j} + \beta_{Ij} I_i + \beta_{Fj} F_i + \beta_{IFj} I_i \cdot F_i + \beta_{Cj} C_i, \quad j = 0, 1, 2; \ i = 1, \dots, N_{housing\ units} \qquad (3)$$

When testing the moderation by existing immigrant populations, it is important to keep in mind that these immigrants were not exogenously assigned: Previous immigrants could move freely within Germany and presumably self-selected into regions with more favorable conditions, such as a lower risk of discrimination. At this point, a main advantage of this two-wave field experiment comes into play: We use the 1st wave of the experiment to estimate and control the baseline level of discrimination in different federal states (or other regional units). By including federal state dummy variables as fixed effects in multivariable analyses, we intend to net out all time-invariant characteristics of federal states, including the baseline discrimination rates that exist there. Furthermore, we controlled for the same federal, regional, advertisement and county-level variables as laid out before. Possible further threats to our identification strategy are discussed in the main text.

S2. Main Results

S2.1. Gross and Net Discrimination Rates and Change by Refugee Immigration

Table S1 reports response rates and gross and net discrimination rates resulting from the paired applicants by wave, underlying Figure 4 in the main text.

Table S1. Detailed results for the discrimination rates presented in Figure 4

Wave 1				Wave 2			
Number of tested housing units: N = 2,389				Number of tested housing units: N = 2,410			
Overall response rate: 58.3 %				Overall response rate: 58.4 %			
Response rate German applicant: 63.5 %				Response rate German applicant: 63.5 %			
Response rate Turkish applicant: 53.1 %				Response rate Turkish applicant: 53.4 %			
		Turkish Applicant				**Turkish Applicant**	
		Response	No response			Response	No response
German Applicant	Response	***Equal treatm.*** (j=0) 48.5% (N=1,159)	***Gross discr. T*** (j=1) 14.9% (N=357)	**German Applicant**	Response	***Equal treatm.*** (j=0) 49.8% (N=1,201)	***Gross discr. T*** (j=1) 13.7% (N=329)
	No response	***Gross discr. G*** (j=2) 4.6% (N=109)	***Equal treatm.*** (j=0) 32.0% (N=764)		No response	***Gross discr. G*** (j=2) 3.5% (N=85)	***Equal treatm.*** (j=0) 33.0% (N=795)
Net discrimination T: 14.9% - 4.6% = 10.3pp				Net discrimination T: 13.7% - 3.5% = 10.2pp			
McNemar's χ^2 (1) = 132.0, p < 0.001				McNemar's χ^2 (1) = 143.8, p < 0.001			

Notes: pp = percentage points.

To see whether there was a significant change in discrimination rates between waves, we estimated multinomial logistic regressions as described in Section S1.2., netting out possible shifts in the composition of federal states, as well as other control variables. The results are shown in Figure S2. The left panel shows average marginal effects of the experiment in the 2nd wave (compared to the 1st wave) on the gross discrimination rates of Turks and Germans in percentage points. Both effects are negative and very similar in magnitude, meaning that there was a parallel decline in the gross discrimination rates of both ethnicities. Both effects fail to reach statistical significance (p > 0.05, as shown by the 95% confidence intervals intersecting with the null line). This parallel decline in discrimination rates suggests that there was no change in the net discrimination rate. This conclusion is supported by a test for the equivalence of the two regression coefficients: the wave effect on the gross discrimination of Turks, and the wave effect on the gross discrimination of Germans. The null hypothesis of no difference cannot be rejected ($\chi^2(1) = 0.32$; $p = 0.570$).

The result of a stable net discrimination also holds for the metric treatment variable, defined by the number of refugees admitted per 100 inhabitants in a given county. This number ranged from nearly 0 up to 4 immigrating refugees per 100 inhabitants for the two waves of this experiment, when excluding the 1% of counties with the highest refugee immigration to obtain outlier-resistant estimates. (Four counties received between 4 and 8 refugees per 100 inhabitants, probably due to hosting large initial reception centers; we exclude these extreme counties from our analyses.) Whereas in 2014 (i.e., before the refugee crisis) the average refugee immigration was only 0.16 per county, the immigration rate rose to 0.75 in 2015.[9] Figure S2 shows that neither of the two gross discrimination rates were significantly affected by the magnitude of refugee immigration (see the regression results in the middle panel in Figure S2). Finally, we also tested the relative increase in refugee numbers per capita compared to the previous year (right panel in Figure S2; the 4 most extreme counties excluded as well). Again, the fairly parallel decline in the two gross discrimination rates means that net discrimination against Turks remained stable. As a further robustness check, we restricted the analysis sample to the 2nd wave for the two metric treatments. Again, the effects are not statistically significant.

[9] We approximate the amount of immigration to counties for the 1st wave (before the refugee crisis) with data on share of refugee immigration on the total population during 2014; and the amount of immigration for the 2nd wave with data on share of refugee immigration during 2015. As can be seen in Figure 2 in the main text, these numbers should be good proxies for the different amount of immigration before each wave of this experiment. Unfortunately, more fine-grained numbers on a monthly/daily basis about immigration to a county are not available.

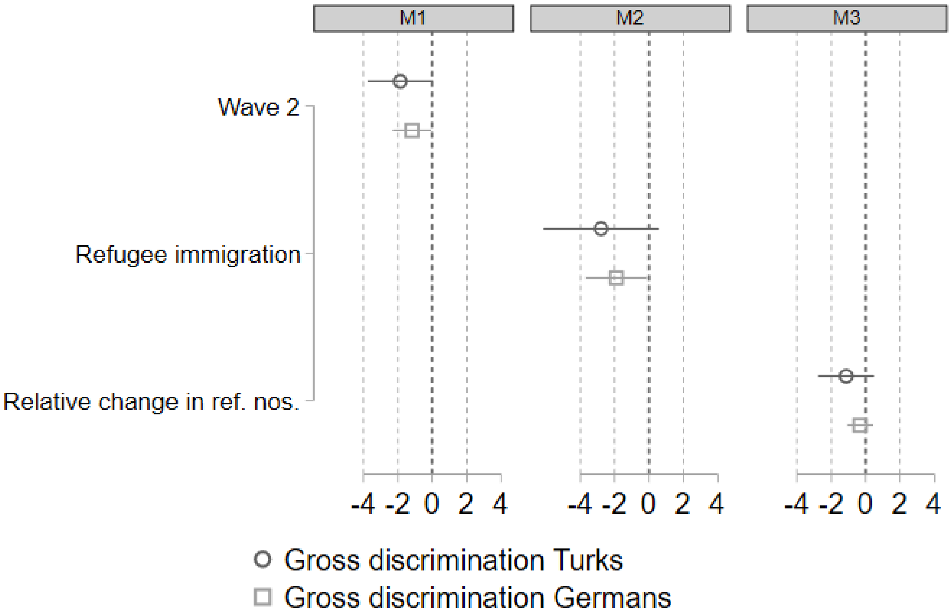


Figure S2. Effects of wave and size of refugee immigration on gross discrimination rates. Results of separate regressions using wave (left panel) as the predictor; the amount of immigration per county (number of newly arrived refugees per 100 inhabitants; middle panel), and the relative size (%-change) of refugee immigration per capita compared to the previous year (right panel: "relative change in refugee numbers"). We report average marginal effects with 95% confidence intervals in percentage points, estimated by multinomial regressions. The sample consists of at least 2,339 tested housing units per wave.

That the decline in gross discrimination rates of Turks and Germans was parallel (i.e., did not significantly differ) was again consistently confirmed by χ^2-tests (testing the null hypothesis that the coefficients for Turks and Germans are equal). For the refugee immigration per capita, the test statistic is: $\chi^2(1) = 0.19$; $p = 0.662$; for the %-change in refugee numbers: $\chi^2(1) = 0.74$; $p = 0.389$. Note that these substantive conclusions remain unchanged when including the four extreme counties with very strong (relative) immigration.

S2.2. Treatment Effect Heterogeneity – Regions with Varying Levels of Foreigners

Table S2 reports the net discrimination rates underlying Figure 5 in the main text, together with the gross discrimination rates and descriptive statistics on response patterns. The share of foreigners per county in 2015 indicates the size of previous immigrant populations and, thus, the extent to which counties were already accustomed to immigration before the refugee crisis. The table is structured in four panels, sorted by the quartile of the share of foreigners (low to high). The results on the left (right) report the response patterns and discrimination levels observed in the 1st (2nd) wave. Generally, net discrimination of Turks was higher in counties with a low share of foreigners; but this pattern did not change across the timing of the two waves (in all panels, net discrimination is similar across the two waves). This result is supported by multivariable analyses. We run multinomial regressions analogously and use the share of foreigners per county as a predictor for the level of gross discrimination. These analyses confirm that there was no significant change between waves. Therefore, although cross-sectional analyses reveal an association between the share of foreigners and the level of discrimination, this association does not seem to be causal.

Table S2. Detailed results for the net discrimination rates presented in Figure 5

Housing units located in counties with lowest share of foreigners (1st quartile: 1.0%-6.0%)

Wave 1

Number of tested housing units: *N* = 575

Overall response rate: 62.4%

Response rate German applicant: 68.2%

Response rate Turkish applicant: 56.7%

		Turkish Applicant	
		Response	No response
German Applicant	Response	***Equal treatm.*** (*j*=0) 53.0% (*N*=305)	***Gross discr. T*** (*j*=1) 15.1% (*N*=87)
	No response	***Gross discr. G*** (*j*=2) 3.7% (*N*=21)	***Equal treatm.*** (*j*=0) 28.2% (*N*=162)

Net discrimination T: 15.1% - 3.7% = 11.4pp

McNemar’s χ^2 (1) = 40.3, $p < 0.001$

Wave 2

Number of tested housing units: *N* = 649

Overall response rate: 60.1%

Response rate German applicant: 66.6%

Response rate Turkish applicant: 53.6%

		Turkish Applicant	
		Response	No response
German Applicant	Response	***Equal treatm.*** (*j*=0) 50.1% (*N*=325)	***Gross discr. T*** (*j*=1) 16.5% (*N*=107)
	No response	***Gross discr. G*** (*j*=2) 3.5% (*N*=23)	***Equal treatm***. (*j*=0) 29.9% (*N*=194)

Net discrimination T: 16.5% - 3.5% = 13.0pp

McNemar’s χ^2 (1) = 54.3, $p < 0.001$

Housing units located in counties with second lowest share of foreigners (2nd quartile: 6.1%-9.8%)

Wave 1

Number of tested housing units: *N* = 592

Overall response rate: 57.7%

Response rate German applicant: 64.9%

Response rate Turkish applicant: 50.5%

		Turkish Applicant	
		Response	No response
German Applicant	Response	***Equal treatm.*** (*j*=0) 47.3% (*N*=280)	***Gross discr. T*** (*j*=1) 17.6% (*N*=104)
	No response	***Gross discr. G*** (*j*=2) 3.2% (*N*=19)	***Equal treatm.*** (*j*=0) 31.9% (*N*=189)

Net discrimination T: 17.6% - 3.2% = 14.4pp

McNemar’s χ^2 (1) = 58.7, $p < 0.001$

Wave 2

Number of tested housing units: *N* = 614

Overall response rate: 61.4%

Response rate German applicant: 66.8%

Response rate Turkish applicant: 56.0%

		Turkish Applicant	
		Response	No response
German Applicant	Response	***Equal treatm.*** (*j*=0) 52.3% (*N*=321)	***Gross discr. T*** (*j*=1) 14.5% (*N*=89)
	No response	***Gross discr. G*** (*j*=2) 3.8% (*N*=23)	***Equal treatm.*** (*j*=0) 29.5% (*N*=181)

Net discrimination T: 14.5% - 3.8% = 10.7pp

McNemar’s χ^2 (1) = 38.9, $p < 0.001$

Housing units located in counties with second highest share of foreigners (3rd quartile: 9.9%-14.3%)

Wave 1

Number of tested housing units: *N* = 704

Overall response rate: 58.6%

Response rate German applicant: 62.9%

Response rate Turkish applicant: 54.3%

		Turkish Applicant	
		Response	No response
German Applicant	Response	***Equal treatm.*** (*j*=0) 47.9% (*N*=337)	***Gross discr. T*** (*j*=1) 15.1% (*N*=106)
	No response	***Gross discr. G*** (*j*=2) 6.4% (*N*=45)	***Equal treatm.*** (*j*=0) 30.7% (*N*=216)

Net discrimination T: 15.1% - 6.4% = 8.7pp

McNemar's χ^2 (1) = 24.6, $p < 0.001$

Wave 2

Number of tested housing units: *N* = 675

Overall response rate: 56.9%

Response rate German applicant: 60.9%

Response rate Turkish applicant: 52.9%

		Turkish Applicant	
		Response	No response
German Applicant	Response	***Equal treatm.*** (*j*=0) 49.3% (*N*=333)	***Gross discr. T*** (*j*=1) 11.6% (*N*=78)
	No response	***Gross discr. G*** (*j*=2) 3.6% (*N*=24)	***Equal treatm.*** (*j*=0) 35.6% (*N*=240)

Net discrimination T: 11.6% - 3.6% = 8.0pp

McNemar's χ^2 (1) = 28.6, $p < 0.001$

Housing units located in counties with highest share of foreigners (4th quartile: 14.4%-32.3%)

Wave 1

Number of tested housing units: *N* = 518

Overall response rate: 53.9%

Response rate German applicant: 57.3%

Response rate Turkish applicant: 50.4%

		Turkish Applicant	
		Response	No response
German Applicant	Response	***Equal treatm.*** (*j*=0) 45.8% (*N*=237)	***Gross discr. T*** (*j*=1) 11.6% (*N*=60)
	No response	***Gross discr. G*** (*j*=2) 4.6% (*N*=24)	***Equal treatm.*** (*j*=0) 38.0% (*N*=197)

Net discrimination T: 11.6% - 4.6% = 7.0pp

McNemar's χ^2 (1) = 15.4, $p < 0.001$

Wave 2

Number of tested housing units: *N* = 472

Overall response rate: 54.4%

Response rate German applicant: 58.7%

Response rate Turkish applicant: 50.2%

		Turkish Applicant	
		Response	No response
German Applicant	Response	***Equal treatm.*** (*j*=0) 47.0% (*N*=222)	***Gross discr. T*** (*j*=1) 11.7% (*N*=55)
	No response	***Gross discr. G*** (*j*=2) 3.2% (*N*=15)	***Equal treatm.*** (*j*=0) 38.1% (*N*=180)

Net discrimination T: 11.7% - 3.2% = 8.5pp

McNemar's χ^2 (1) = 22.9, $p < 0.001$

Notes. pp = percentage points.

S3. Robustness and Balance Checks

We conducted several checks to validate our results.

S3.1. Robustness Checks

Alternative Treatment: Male Refugees and Registered Asylum Seekers

As alternative treatment variables, we used information on the number of male refugees and on registered asylum seekers. First, a higher share of male refugees may have a stronger impact on a change in discriminatory behavior, as experimental evidence shows that respondents are less likely to trust immigrant men than immigrant women (Gereke, Schaub and Baldassarri 2020). However, the multinomial logit estimations show very similar results to the original treatment that includes both genders (Figure S3).

Second, we used the number of newly registered asylum seekers (German: "Schutzsuchende") per 100 inhabitants as a further alternative treatment variable that encompasses not only refugees currently involved in an asylum procedure, but also those with protection status.[10] This treatment variable also shows very similar results (Figure S3).

[10] The data originate from the Central Register of Foreigners (German: "Ausländerzentralregister").

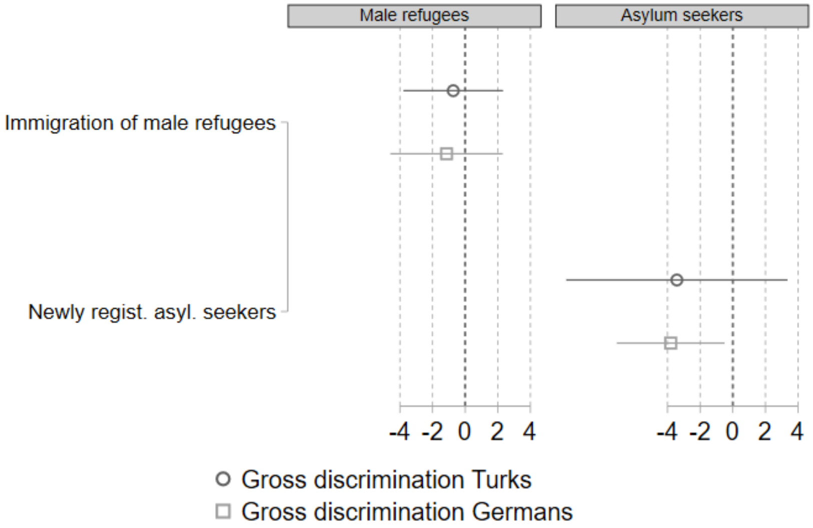


Figure S3. Effects of the immigration of male refugees (per 100 inhabitants) and of newly registered asylum seekers (per 100 inhabitants) on gross discrimination. We report average marginal effects with 95% confidence intervals in percentage points, estimated by multinomial regressions. The sample consists of at least 2,336 tested housing units per wave.

Alternative Treatment: Walking Distance to Refugee Reception Centers

Our main analysis relies on county-level data about the presence of new refugees. To test whether changes in discrimination occurred on a more fine-grained spatial level, we conducted an additional analysis at the neighborhood level using multinomial logit regressions.

For this, we use the exact geographical location of refugee reception centers (RRC) in five federal states and the walking distance from the sampled housing units as calculated by Google Maps. RRCs are run by the federal states, and data is therefore provided by state officials. RRCs are the first accommodations for refugees where they file their asylum applications. In the second half of 2015, the large number of arrivals made it necessary to install several emergency accommodations, for instance, in high schools, community halls, hotels, and hostels. These alternative accommodation locations are also included in our data. Other types of refugee accommodation, such as community housing, are managed by the municipalities and are thus not included in our data. However, Berlin is an exception – here, we have location data for all types of refugee accommodations.

Table S3. Number of refugee reception centers (RRC) per federal state and wave

	Wave 1	**Wave 2**
Baden-Wuerttemberg	18	43
Berlin	62	80
Hesse	2	50
Lower Saxony	3	32
Saxony	7	38
Total	92	243

Notes: For the other 11 federal states, the exact geographic locations were not provided by the federal state ministries due to data protection reasons.

Only housing units with an exact location (street and house number) are included in the analysis sample. Further, to analyze effects in the immediate neighborhood, the sample is restricted to rental units near an RRC (geodesic distance < 10km). The final sample comprises 604 housing units (1st wave: 233; 2nd wave: 371).

When analyzing the effects of the distance to refugee shelters, one must consider that the placement of shelters at the local level within counties was probably not entirely random, as the quota system for refugee distribution applied only to the county or community level. The administrators responsible for installing RRCs within counties or communities certainly considered factors such as the acceptability of refugees in different neighborhoods, available vacancies, and rent levels (Hennig 2021). These variables could be correlated with tastes for discrimination or economic motives underlying statistical discrimination. However, we can bypass this endogeneity issue by using the 1st wave of the experiment to statistically control for the baseline level of discrimination in different neighborhoods (caused by these or other time-invariant variables). At the time of the 1st wave, the refugee crisis and, thus, the locations of future refugee shelters were not yet foreseeable and could therefore not yet have influenced the level of discrimination. This allows us to identify the net effect of refugee shelter locations.

Multinomial logit regression models estimate the effect of the walking distance and wave interaction effects on the discrimination of Turkish and German applicants. Results of the first model show that neither the walking distance nor the interaction of the walking distance with the 2nd wave significantly affects discrimination (Figure S4). In other words, even during the refugee crisis, suppliers of rental units in close proximity to RRCs did not discriminate more strongly than before the start of the crisis.

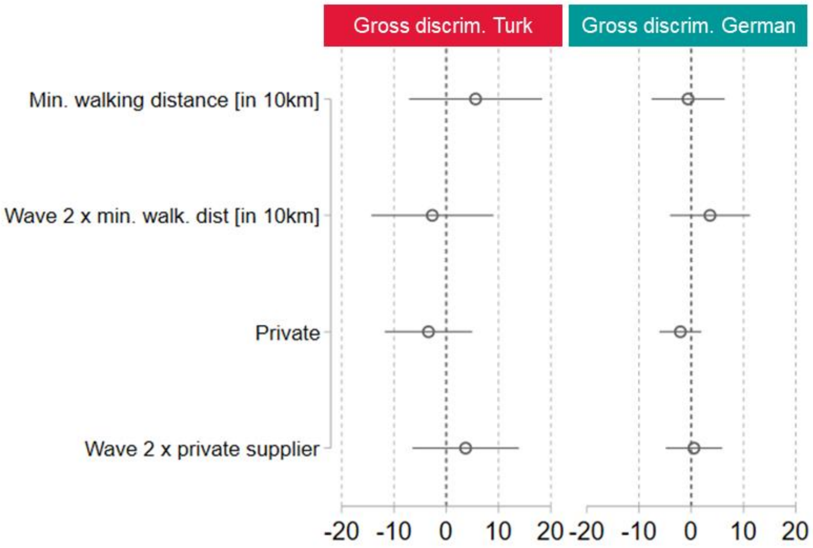


Figure S4. Multinomial logit – Minimum walking distance to closest RRC on discrimination. Average marginal effects in percentage points with 95% confidence intervals are shown. The sample consists of at least 233 tested housing units per wave.

In conclusion, the fine-grained spatial analysis confirms our main results: The close proximity of a rental housing unit to an RRC does not seem to affect the suppliers' tendency to discriminate.

Alternative Outcomes: Invitation to Viewings and Response Times

Following standard procedures, the main outcome variable measuring discrimination is response vs. non-response. We used the response content and response time as alternative outcomes for robustness analyses. Almost all responses were invitations to view the rental unit; only few responses included offers to view other units (e.g., because the unit was no longer available), and some responses were more difficult to categorize (e.g., requests to call the supplier). We coded all explicit invitations to a viewing as a positive response. All other responses and non-response were coded as a negative response. Discrimination rates based on this variable differed only slightly from our main treatment. For example, the gross discrimination of Turks observed in the 1st wave based on invitations was 14.9 percent, the gross discrimination of Germans 4.6 percent, and the net discrimination 10.3 percentage points, which is the same net discrimination rate as we observed with our main outcome.

Time to respond was used as a metric measurement of different treatment, and discrimination was here measured by mean comparisons between applicants of different ethnicity. For analysis, we used Cox regressions that allow to include right-censored response times. Also, the analyses with this alternative outcome, response time, confirmed the conclusion of our main analysis: The level of discrimination did not change between waves (analyses are available on request).

Non-linear Effects

For the metric treatment variable (i.e., the magnitude of immigration to different counties), we also analyzed possible non-linear relationships with the levels of gross and net discrimination. Figure S5 shows a local polynomial smoothing of the levels of discrimination observed in the 2nd wave across different numbers of refugees that counties received per 100 inhabitants. The 5 percent of counties with the lowest and highest immigration rates have been excluded to obtain outlier-resistant estimates. This nonparametric approach does not require any assumptions about the functional form of the effect. Therefore, possible non-linear relationships can be identified. If the theory of tipping points were supported, this would result in non-linear effects (Galster 2014). However, neither the two gross discrimination rates nor their difference (i.e., net discrimination rate) were affected by the magnitude of refugee immigration. These results also hold when using the relative increase in the proportion of refugees hosted in counties as the treatment variable (analyses available on request). More in-depth analyses with comparisons across waves only show that in counties that received 2 to 8 asylum seekers per 100 residents in 2015, discrimination against Turkish applicants slightly increased compared to 2014. However, due to the small number of cases observed in these extreme counties (N = 46 housing units in the 2nd wave), these results do not reach a statistically significant level and should be interpreted cautiously. Thus, we again conclude that discrimination against Turkish relative to German applicants has not changed substantially during the refugee crisis.

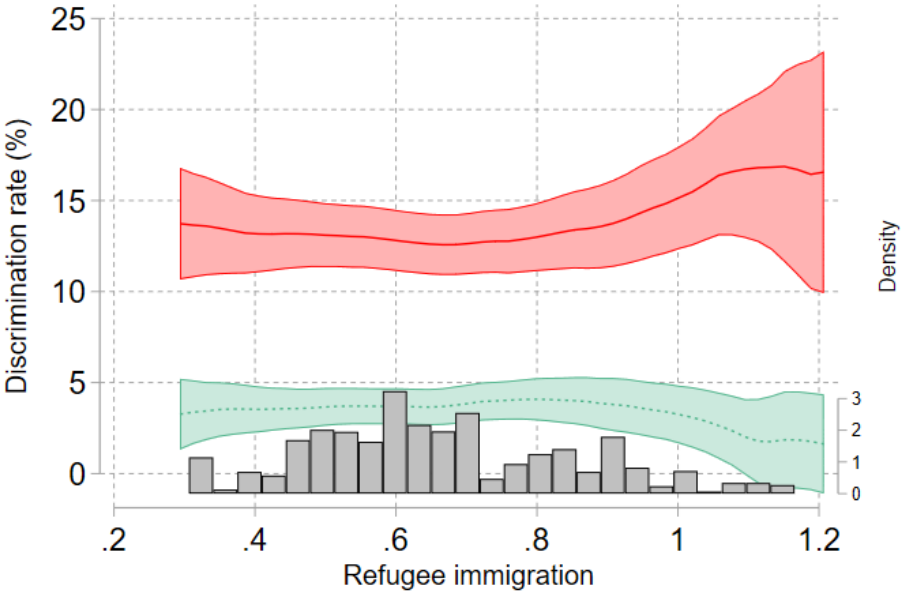


Figure S5. Gross discrimination rates of Turkish (red, solid line) and German (green, dashed line) applicants in the 2nd wave by refugee immigration per county in 2015 (*N* refugees received per 100 inhabitants). Net discrimination is the gap between these two gross discrimination rates. The graph shows local polynomial smoothing with 95% confidence intervals based on 325 counties (5% of counties with smallest and highest immigration excluded to achieve stable results; for these counties, with only few observations, inflated confidence intervals do not allow for clear results). The histogram at the bottom summarizes the density distribution of counties with different sizes of refugee immigration. The sample consists of 2,248 tested housing units.

Further Robustness Analyses for the Main Identification Strategy

We also tested whether our results still hold when restricting the analysis sample to housing units located in Western Germany, or advertised by private or commercial suppliers (real estate agencies) only. In addition, we used the extensive information on the housing market collected to ensure that our results do not suffer from "length bias" or other sample restrictions.[11] For detailed analyses of these possible method effects, see Auspurg, Schneck and Thiel (2020).

[11] A length bias could occur if listings advertised over a longer period are more likely to be included in our sample of tested housing units.

Alternative Identification Technique: Instrumental Variable Approach

As an alternative identification method for the effect of refugee immigration on discrimination, we used the refugee crisis as an instrumental variable (IV). This variable (indicated with W in equation 4) equals 0 in the 1st and 1 in the 2nd wave of our field experiment. For a two-stage-least-square estimation, two assumptions have to be met. First, according to the *relevance condition*, the IV must be correlated with the independent variable of interest (registered immigrating refugees, I). Second, the *exclusion restriction* requires that an IV must be uncorrelated with the error terms $u_{iG}\ resp. u_{iT}$ (see equations 5a,b), which represent the unexplained discrimination rates that remain when netting out the effect of the share of refugees $\hat{I}$. The relevance condition is clearly met, as the first regression estimation based on the Two-Stage-Least-Squares (2SLS) approach (equation 4) shows a substantial effect, and the Pearson correlation is substantial with $r(N = 4{,}799) = 0.58$, $p < 0.001$. Therefore, our IV is a strong instrument (for more details on the issues of weak instruments, see Andrews, Stock and Sun 2019).

$$I_i = \beta_0 + \beta_W W_i + u_i, \qquad i =, \dots, N_{housing\ units} \tag{4}$$

The second step of 2SLS identifies the actual effect of interest (see equation 5a and 5b). Here, the expected value of the share of immigrating refugees in each county from equation 4 ($\hat{I}$) is used as an IV on discrimination. Since we use a linear regression model, the multinomial outcome was split into two dichotomous outcomes: the discrimination of the Turkish (j = 1, vs. j = 0 & j = 2, D_T) and German (j = 2, vs. j = 0 & j = 1, D_G) applicant.

$$D_{iT} = \beta_{0T} + \beta_{\hat{I}T}\widehat{I_\imath} + u_{iT} \tag{5a}$$

$$D_{iG} = \beta_{0G} + \beta_{\hat{I}G}\hat{I}_i + u_{iG} \tag{5b}$$

If the exclusion restriction holds, the coefficients $\beta_{\hat{I}T}$ and $\beta_{\hat{I}G}$ are unbiased causal estimates of the effect of the share of refugees in a county on the gross discrimination of Turks or Germans. Since our instrument is exogenous due to a natural experiment setting, one can expect that the exclusion restriction is fulfilled (see the discussion in the main text). To prevent a violation of the heteroskedasticity assumption, which may yield biased standard errors, we used cluster-robust standard errors on the county-level (Rogers 1993). In order to account for effect heterogeneity across subgroups, we also run the IV models separately for regions (East/West), county size (city/rural), and type of housing supplier (private/corporate).

The results confirm our previous findings. Neither the discrimination of the Turkish applicant nor of the German applicant changed significantly over time (Figure S6). Similarly, examining potential heterogeneous effects across subgroups showed consistent but non-significant results. This finding shows that there are neither regional differences (East/West) nor between urban and rural areas, nor between private and corporate housing suppliers. Similar to our main findings, there is no decisive heterogeneity for different subgroups.

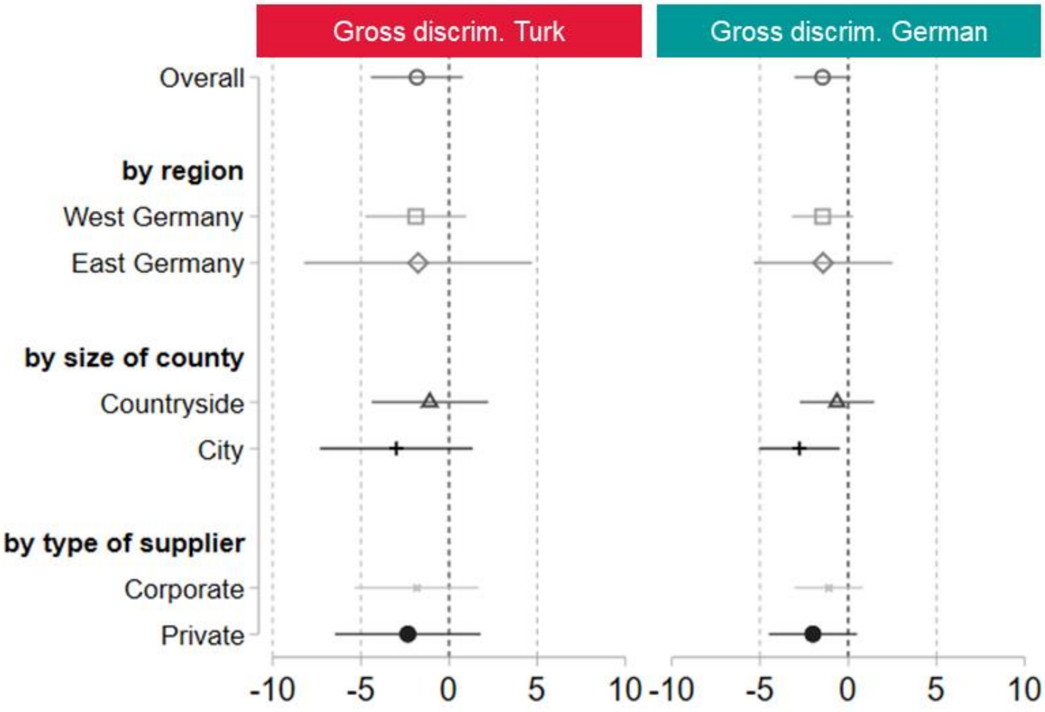


Figure S6. Results of 2SLS estimation (instrumental variable: wave) using linear probability models. Coefficients with 95% confidence intervals are shown. Each coefficient represents a separate analysis. The scale shows the difference in percentage points. *N* = 4,799.

S3.2. Balance Checks

Random Assignment and Balance of Field Experimental Factors

We carefully checked whether the randomization of our experiment worked on several important dimensions to achieve maximum internal validity. First, to identify the total effects of ethnicity and thus the level of ethnic discrimination, it is important that ethnicity is not correlated with any of the other applicant characteristics that may affect housing suppliers' replies. Second, the desired balance in the levels of treatment variables to achieve a maximum (independent) variance in these variables was also successfully realized (i.e., all levels of an applicant characteristic occurred with about the same frequency). Third, it is important that characteristics of the two applicants applying to the same unit are not inter-correlated; otherwise, there might be idiosyncratic effects due to specific (non-random) pairings of applicants. Table S4 shows that this was also achieved: All correlations across characteristics of applicants show negligible effect sizes close to zero. We also ensured that the different text versions used to conceal the nature of the experiment (i.e., different salutations or orders of text phrases in the e-mails) did not evoke any idiosyncratic response patterns. The maximum correlation of a text version with the observed response pattern (i.e., observed level of discrimination) was r = 0.025, p = 0.17 (cf. Auspurg, Schneck and Thiel 2020: 5). This confirms that the randomization of the text versions worked.

Table S4. Descriptive statistics on realized experimental design by ethnicity and χ^2-Test/*t*-test for statistical group-difference (*p*-value).

	Turkish applicant	**German applicant**	**χ^2-test**	***t*-test**
Occupational level			*p* = 0.750	
no information	33.9%	33.8%		
low (vocational training)	33.3%	32.7%		
high (university degree)	32.8%	33.5%		
Employment status			*p* = 0.108	
no information	25.6%	24.0%		
employed	24.9%	25.0%		
self-employed	24.3%	26.3%		
public service	25.1%	24.7%		
Family status			*p* = 0.553	
single	38.9%	38.40%		
couple	30.0%	31.0%		
family	31.1%	30.6%		
Mean income (in €)	1,678	1,715		*p* = 0.272

Notes. Applicant income was included in a random subsample of e-mails as additional information.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (two-tailed test).

We further tested whether any names provoked an idiosyncratic effect (see Figure S7). In each ethnicity group, only one out of 30 names showed a response rate that significantly differed from the mean (up or down; 5% significance level). This corresponds to a rate that is expected to occur by chance (due to the "alpha error" in significance tests).

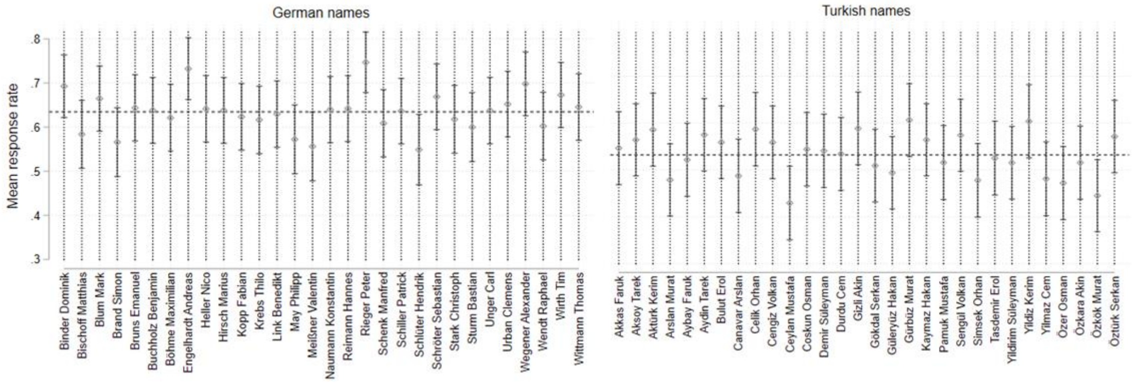


Figure S7. Response rates by ethnicity and names. Coefficients from multivariable regressions including only the names as dummy variables with 95% confidence intervals. Both waves pooled (*N* = 4,799 for both Turks and Germans).

Stability (Balance) of Experimental Variables over Time

We tested whether the means or distributions of the experimental variables differed by wave (see Table S5). If the randomization and/or repetition of our experiments across waves was successful, there would be no differences. Table S5 proves that there were no statistically significant changes between waves (*p* > 0.05). In addition, the variables are rather evenly distributed within each wave, indicating a high level balance.

Table S5. Descriptive statistics on realized experimental design by wave and χ²-Test/*t*-test for statistical group-difference (*p*-value).

	Turkish applicant				**German applicant**			
	Wave 1	**Wave 2**	**χ²-test**	***t*-test**	**Wave 1**	**Wave 2**	**χ²-test**	***t*-test**
Occupational level			*p* = 0.744				*p* = 0.705	
no information	33.5%	34.2%			34.3%	33.6%		
low (voc. train.)	33.2%	32.2%			32.7%	33.8%		
high (uni degree)	33.3%	33.7%			33.0%	32.6%		
Employment status			*p* = 0.732				*p* = 0.759	
no information	24.2%	23.9%			25.4%	25.9%		
employed	24.9%	25.0%			25.0%	24.9%		
self-employed	26.8%	25.7%			23.9%	24.7%		
public service	24.1%	25.3%			25.7%	24.5%		
Family status			*p* = 0.189				*p* = 0.167	
single	37.7%	39.1%			39.8%	38.0%		
couple	32.2%	29.8%			28.8%	31.2%		
family	30.1%	31.1%			31.4%	30.8%		
Mean income (in €)	1,780	1,773		*p* = 0.880	1,719	1,763		*p* = 0.372

Notes. For the mean income, see notes for Table S4.
* *p* < 0.05; ** *p* < 0.01; *** *p* < 0.001 (two-tailed test).

Maps of Refugees and Foreigners per County

By law, refugees are allocated to the 16 federal states according to a strict quota system. Within a federal state, the allocation to the various counties follows different rules. Therefore, it is necessary to confirm whether the assumption for the identification of treatment heterogeneity, that the resulting distribution at the county level was random, is met. While the distribution of foreigners (i.e., without German citizenship) was highly geographically clustered in 2015, supporting the thesis that there was a strong self-selection into specific regions (Figure S8, right), the distribution of refugees during the crisis was much more random (see Figure S8, left). Statistics and spatial regressions verify that the regional influx of refugees was not auto-correlated with any observable state or county characteristic (see Table S6).

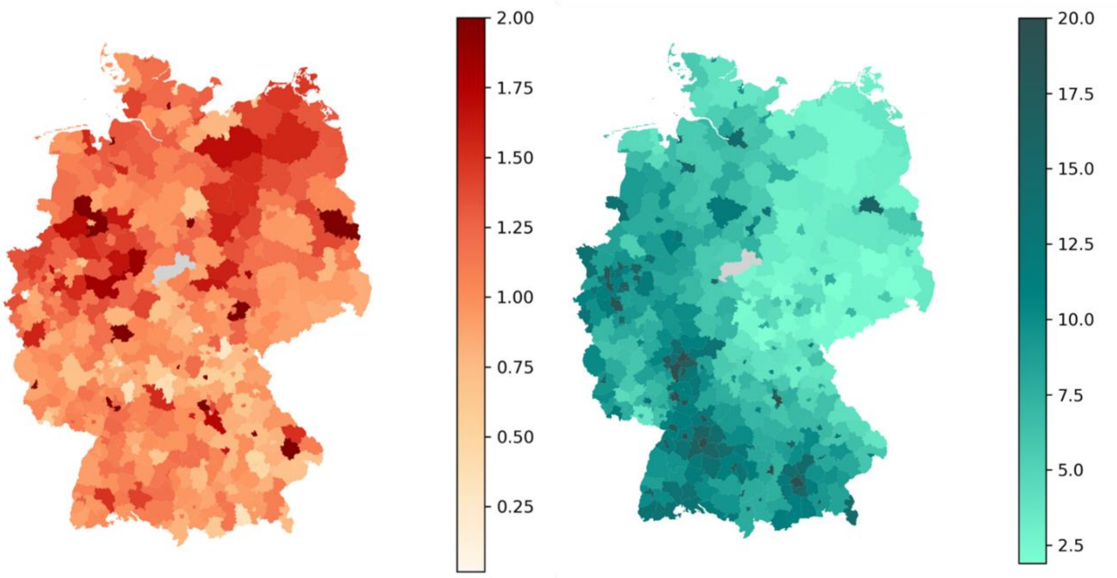


Figure S8. Map of refugees per county in 2015 (left, outliers >2 recoded to 2) and map of foreigners in 2015 (right, outliers >20 recoded to 20). Both per 100 inhabitants. Foreigners are defined as those without a German citizenship. Data: Federal Ministry of Interior. Own illustration.

Balance Check of Control and Treatment Group

This section tests whether our main treatment of interest, the wave and amount of influx of refugees, was exogenous to other variables that might have affected discrimination rates. First and most importantly, there should be no correlation with the ethnicity of the applicants. This is *per se* achieved by our paired experimental design. Second, other applicant characteristics should also be perfectly balanced (randomized) across waves, as tested in the previous section. Finally, for our binary treatment (1st vs. 2nd wave) to show internal validity, it is also important that other characteristics of the housing market that might affect differential treatment of Turkish vs. German applicants did not change over time, and thus did not confound the influx of refugees between waves. Since the latter is out of our experimental control, we might at least identify possible confounders to control in multivariable analyses. As characteristics that are stable over time can be eliminated by regression models with federal-state fixed effects, we included federal-state dummy variables in multivariable analyses

To test whether there was balance in characteristics of the housing market and contextual variables between the two waves of our experiment (i.e., to test the *ignorability* assumption), we used LPM models of our wave dummy as the dependent variable and apartment characteristics (private supplier (yes/no), number of rooms, rent per sqm), regional characteristics (county fixed effects, city) as well as county-level variables (share of foreigners (2014), unemployment rate (2014, 2015), population density (2014, 2015), GDP per person employed (2014, 2015), the vacancy rate (2011, census data) as well as the voter share for the green party in the previous federal election (2013) in bivariate models. For the apartment characteristics, we observe a significant increase in the proportion of private suppliers due to a legal change in the German housing market (see Figure S9).[12] We cover this possible confounding factor by using this type of supplier as a control variable. In addition, we run robustness analyses with this variable as sub-grouping variable in regression models. No statistically significant differences are observed for the number of rooms and the rent per square meter. Also no statistically significant compositional differences could be observed for the regional or county-level characteristics, with the only exception that the apartments tested in the 2nd wave were located in counties with a slightly lower unemployment rate and a slightly lower share of green party voters compared to those tested in the 1st wave. Whereas the decrease in unemployment could imply lower threats by competition and therefore, a lower discrimination of Turks, the decrease of apartments located in regions with a high share of green voters (a party supporting immigration) may point to a slightly less migrant-friendly population in the 2nd wave, which would result in an increase in discrimination of Turks. We therefore include these two variables also in our robustness analyses. The overall test of county fixed-effects shows no imbalances between our two waves: overall, all counties are represented to a similar extent in both waves ($F(387, 4411) = 0.95$, $p = 0.754$, omitted in Figure S9 due to space constraints). These results were also confirmed in multivariable regressions.

[12] Since June, 1 2015, the so-called “buyer pays principle” has applied to the renting of flats. According to this, the person who has commissioned an agency—usually the housing supplier—pays the agency. This legislative change led to a small increase in flats rented out by private providers (i.e. without using estate agents), as they can no longer shift the costs for this onto the renters.

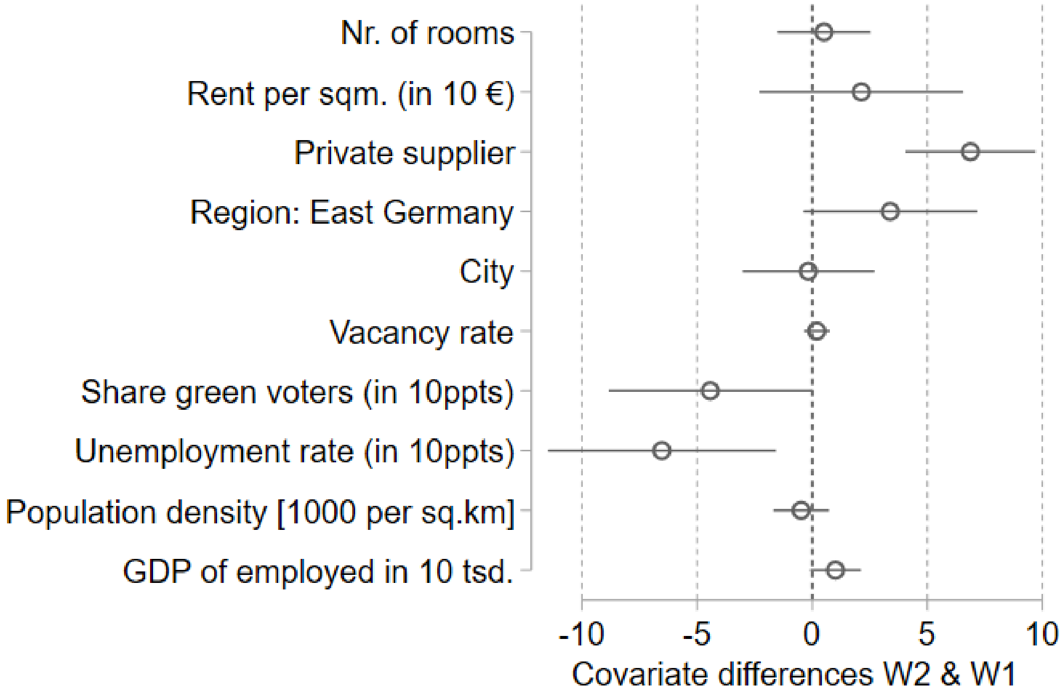


Figure S9. Effects of apartment, regional and context characteristics on the probability that an apartment with these characteristics was tested in the 2nd instead of 1st wave of the field experiment in bivariate regressions. Coefficients with 95% confidence intervals are shown. Each coefficient represents a separate bivariate analysis (LPM-model). The scale shows the effects in percentage points. Non-significant coefficients indicate balance across both waves; while statistically significant coefficients indicate a shift in the sample composition across waves. At least N = 4725 apartments (discrepancies to the N = 4,799 tested apartments arise from few missing values for variables on the regional level, especially for the vacancy rates).

Independence of Refugee Immigration from Changes in Housing Market

Our second research question, examining the treatment heterogeneity of the effect of the refugee crisis in a natural experiment, relies on the assumption that the assignment treatment and control group must be independent of other regional characteristics that may confound the interaction of the treatment with the (observational) share of foreigners in 2014. To test for such a random allocation, we calculated each regional characteristic's Pearson correlation coefficient (r) with the proportion of refugees for both waves on the county level (at least N = 351 included in the experiment in both waves). As expected, county GDP before refugee immigration (pretreatment) is positively correlated, but only with a marginally statistically significant effect size (p < 0.1). Furthermore, counties with a higher unemployment rate received a larger share of refugees before the refugee crisis. After the onset of the refugee crisis, no distinct pattern of refuge allocation could be observed. Therefore we are confident that the effect of previous exposure to immigration on discrimination is independent of regional characteristics.

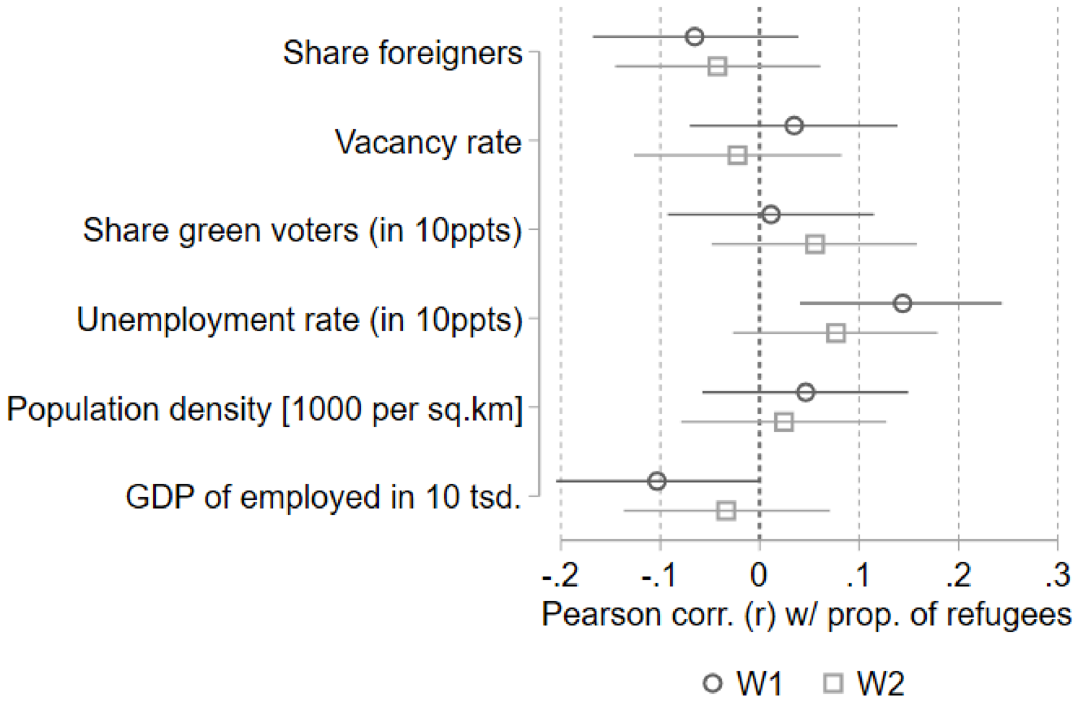


Figure S10. Pearson correlation coefficient (r) of context characteristics with the assigned share of refugees before (1st wave) and after the onset of the refugee crisis (2nd wave). Coefficients with 95% confidence intervals are shown. Each correlation coefficient represents a separate analysis. All analysis on the county level, $N \geq 351$.

In addition, a check for spatial dependency did not show a statistically significant spatial autocorrelation of the refugee distribution (Table S6). We determined Moran's I, one of the central indicators for spatial dependence.[13] The indicator ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation). Moran's I is close to zero for the refugee variable and not statistically significant (see Table S6). This confirms that there is no spatial clustering of refugees at the county level. In contrast, the spatial distribution of foreigners is clearly non-random, as the high value of Moran's *I* reveals (for more information on spatial analysis, see Darmofal 2015).

Table S6. Spatial randomization check

	Moran's *I*	***p*-value**
% Refugees in 2015	0.015	0.198
% New refugees in 2015	0.012	0.197
% Foreigners 2015	0.559	0.001**

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (two-tailed test).

[13] Moran's *I* is based on a spatial weights matrix. This matrix specifies for each county-pair if two counties are neighbors. We used the "queen contiguity" as a definition for the weighting matrix, which defines all adjacent counties as neighbors (common edge or common vertex, reflecting the queen's direction of movements in chess).

SI References

Andrews, Isaiah, James H Stock and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* 11:727-53.

Auspurg, Katrin and Thomas Hinz. 2015. *Factorial Survey Experiments*. Thousand Oaks: SAGE.

Auspurg, Katrin, Andreas Schneck and Fabian Thiel. 2020. "Different Samples, Different Results? How Sampling Techniques Affect the Results of Field Experiments on Ethnic Discrimination." *Research in Social Stratification and Mobility* 65:100444.

Darmofal, David. 2015. *Spatial analysis for the social sciences*: Cambridge University Press.

Gaddis, Michael S. 2017. "How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies ". *Sociological Science* 19(4):469-89.

Galster, George. 2014. "Nonlinear and Threshold Aspects of Neighborhood Effects." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 66(1):117-33.

Gereke, Johanna, Max Schaub and Delia Baldassarri. 2020. "Gendered Discrimination against Immigrants: Experimental Evidence." *Frontiers in Sociology* 5:59.

German Federal Statistical Office - Destatis. 2021. "Migration and Integration - Integration Indicators."

Greene, William H. 2012. *Econometric Analysis*. Boston: Prentice Hall.

Hennig, Jakob. 2021. "Neighborhood Quality and Opposition to Immigration: Evidence from German Refugee Shelters." *Journal of Development Economics* 150:102604.

Phillips, David C. 2019. "Do Comparisons of Fictional Applicants Measure Discrimination When Search Externalities are Present? Evidence from Existing Experiments." *The Economic Journal* 129(621):2240-64.

Rogers, William. 1993. "Quantile Regression Standard Errors." *Stata Technical Bulletin* 2(9).

Vuolo, Mike, Christopher Uggen and Sarah Lageson. 2016. "Statistical Power in Experimental Audit Studies:Cautions and Calculations for Matched Tests With Nominal Outcomes." *Sociological Methods & Research* 45(2):260-303.

Vuolo, Mike, Christopher Uggen and Sarah Lageson. 2018. "To Match or Not to Match? Statistical and Substantive Considerations in Audit Design and Analysis." Pp. 119-40 in *Audit studies: Behind the scenes with theory, method, and nuance*, edited by M. S. Gaddis. Cham: Springer.