Auspurg, Katrin; Schneck, Andreas; Thiel, Fabian

Article — Accepted Manuscript (Postprint)

Different samples, different results? How sampling techniques affect the results of field experiments on ethnic discrimination

Research in Social Stratification and Mobility

Suggested Citation: Auspurg, Katrin; Schneck, Andreas; Thiel, Fabian (2020) : Different samples, different results? How sampling techniques affect the results of field experiments on ethnic discrimination, Research in Social Stratification and Mobility, ISSN 0276-5624, Elsevier, Amsterdam, Vol. forthcoming,
https://doi.org/10.1016/j.rssm.2019.100444 ,
http://www.sciencedirect.com/science/article/pii/S0276562419300575

Different Samples, Different Results?

How Sampling Techniques Affect the Results of Field Experiments on Ethnic Discrimination

Katrin Auspurg[1 2], Andreas Schneck[1], and Fabian Thiel[1]

[1] Department of Sociology, LMU Munich

[2] Corresponding Author: katrin.auspurg@lmu.de

Author accepted manuscript, please cite as:

Auspurg, K., Schneck, A., & Thiel, F. (2020). Different Samples, Different Results? How Sampling Techniques Affect the Results of Field Experiments on Ethnic Discrimination. *Research in Social Stratification and Mobility*, https://doi.org/10.1016/j.rssm.2019.100444.

Abstract

This paper explores a possible sampling bias in field experiments through a unique combination of a large-scale field experiment on ethnic discrimination in the German rental housing market (N = 2,992 tested apartments) and data on the internet housing market where the apartments were sampled from (observation of the whole platform for about one year). Up to now, most field experiments sampled on the level of suppliers (not apartments) and selected the tested suppliers within a short field period (point sampling). This probably led to an over-representation of small suppliers and apartments that had already been advertised for a long time, resulting in an under-representation of large suppliers and new vacancies. We analyze how both issues affect the measurement of ethnic discrimination and its underlying mechanisms. With our case study on the German housing market, we first observed the expected sampling bias: There was a strong over-sampling of small suppliers and apartments already offered for a long time. Second, this bias was found to have only little impact on the descriptive result of substantial discrimination against Turkish applicants. There was only a slight tendency that discrimination was higher for small suppliers and offers advertised for a relatively short time (i.e. new vacancies). Overall, we conclude that evidence on ethnic discrimination and its underlying risk factors is remarkably robust to the used sampling technique. Although there were no indications of a severe bias, our market data illustrate opportunities to test the effects of market conditions that often remained unnoticed.

1. Introduction

This study presents a unique combination of a field experiment on ethnic discrimination with big data on the internet platform where the tested housing units were sampled from. These data inform on the size of suppliers and the time market offers were advertised online: Not only for the sample used in the field experiments, but on all offers advertised on the markt platform. This combination allows us to explore whether standard sampling techniques used in field experiments lower their external validity. As we will argue, these techniques frequently over-sampled small suppliers. When using short field periods, they additionally over-sampled offers with a low success rate (i.e. long advertisement time). This over-representation of units with a long survival time has been termed *length bias* in the statistical literature on observational data (van Es, Klaassen, & Oudshoorn, 2000). To our knowledge, there is so far no literature that explores these sampling issues for field experiments. We will illustrate the resulting biases with a case study on the German rental housing market. However, the main results probably also generalize to other markets, such as the labor or product market, as the sampling strategies tested in our study are standard to field experimental methods in general.

Regarding housing markets, there is large evidence for Western countries that migrants are disadvantaged compared to the majority population. They live, for instance, in apartments with relatively few amenities; they face larger housing costs; and they are also more likely to live in poorer neighborhoods than the majority population (for statistics on Germany: Drever & Clark, 2002; Harrison, Law, & Phillips, 2005). These disparities intersect with other social inequalities, such as poor health, education, and labor market outcomes (Galster, 1992, 1996; Pager, 2008; Turner, Ross, Galster, & Yinger, 2002). Given that, and given an increasing shortage of affordable housing in many Western metropolises with booming labor markets (for Europe: Ball, 2016; for the U.S.: Metcalf, 2018), access to housing is increasingly seen as a crucial factor for determining ethnic minorities' position in the social stratification.

It is thus not surprising that there is substantial interest in the question of the extent to which access to housing is restricted by ethnic discrimination. Starting in the 1960s, there is a large strand of research that used field experiments to answer this question (for two recent meta-analyses: Auspurg, Schneck, & Hinz, 2019b; Flage, 2018). Following the standard procedure in experiments, most scholars focused solely on the internal validity (randomization) of the experiments and ignored how non-representativeness of their selected (housing) units could have impacted their findings. With the possibility of running the field experiments in natural environments, the experiments were simply thought to allow for a sufficiently high amount of external validity. However, when offers of suppliers with a different tendency to discriminate have a distinct probability of being sampled for the experiment, the external validity might be threatened (Bell & Stuart, 2016). In this article, we will discuss in particular two threats to the validity, which are—to the best of our knowledge—missing in the literature on field experiments.

First, while some researchers sampled each housing unit with an equal probability (never minding whether they belong to the same or different suppliers), other researchers sampled each supplier only once. The latter is mainly done for ethical reasons: Sampling each supplier only once means lowering the loss of time and other burdens for the landlords and agencies that are tested in the experiments (for a general discussion of ethical issues in field experiments: Riach & Rich, 2004). However, sampling on the level of suppliers (and not apartments) might also impact the external validity. Compared to the population of all housing vacancies, this will result in an over-representation of vacancies that are offered by small suppliers.

Second, sampling always took place during a limited field period that lasted between one week and several months. In this period, a cross-section of units advertised on the market was sampled. Using such a cross-section induces what has been called a *length bias* in the statistical literature on sampling (van Es, et al., 2000): The probability of being sampled is proportional to the length of time a housing unit stays on the market (is advertised). Thus, units with a low success rate (long survival time on the market platform) are more likely to be sampled than units with a high success rate (low survival time).

Given both issues, the tested units certainly do not represent a random selection from the population of (housing) offers. Small suppliers are mostly private landlords that follow less formal standards than commercial agencies, while units with a low success rate are likely those that are of relatively low quality (given the price) and/or located in deprived areas where it is difficult to find any renters. As we will argue in more detail later on, there are good reasons to believe that this translates to a biased measurement of discrimination. If this is true, the inconsistency of findings on the amount and kind of discrimination (tastes or statistical discrimination) might partly stem from different sampling techniques.

To test our assumptions, we used a novel combination of field experimental data and market data. In 2015, we conducted a large-scale field experiment (e-mail correspondence test) on ethnic discrimination in the rental housing market in Germany. What is specific to our study is that we combined this experiment with rich data on the internet platform where the apartments were advertised. For about one year, we observed the whole internet platform with more than one million advertised housing units. In this study, we analyze 2,992 units tested in Western Germany where we have full information on the size of the supplier (measured by the number of offers advertised during our one-year observation period) and length of the advertisement time before experimental treatment (i.e. we sent our e-mails to apply for the apartment, using Turkish and German identities). To what extent is the sample of our field experiment biased regarding the size of suppliers and/or the length of advertisements? Does this affect the observed discrimination rates? And is there evidence that sampling bias affects not only the level but also the observed mechanisms underlying discrimination, such as the incidence of statistical or taste-based discrimination?

2. State of the Art: Sampling Strategies Used in Field Experiments

Field experiments were developed as a particularly appealing method to measure discrimination in (housing) markets (Pager, 2008).[1] In prior decades, researchers mainly used in-person audits (e.g. Yinger, 1986), where test persons with different ethnic backgrounds apply to the same housing units. With apartments becoming increasingly advertised on the internet, audits have been more and more replaced by correspondence studies where standardized, written e-mail applications are sent out with the names of applicants signaling their ethnicity (for a review of different field experimental methods see e.g. Bertrand & Duflo, 2017). In these experiments, discrimination is typically measured by the differences in the response probabilities to the e-mails sent by minority versus majority applicants. Due to the experimental manipulation of ethnicity, these e-mail experiments provide a high amount of confidence (internal validity) that the ethnicity did cause the variation in the observed outcomes.

Meta-analyses of these experiments document a substantial amount of discrimination for the U.S. and Europe, with a high amount of variation between but also within countries (Auspurg, et al., 2019b; Flage, 2018). Audit studies are seen to suffer from methodological weaknesses that may threaten the internal validity of the results (such as experimenter demand effects and unobserved treatment heterogeneity, see e.g. Heckman, 1998). In the following, we therefore review only the most recent e-mail correspondence tests published between 2010 and 2019 (see Table 1).

In this period, $N = 20$ e-mail correspondence tests covered a wide range of European countries and the U.S., as well as a wide range of tested ethnicities. Overall, this specific subpopulation of more recent tests provided very robust evidence for ethnic discrimination. Compared to the response probability of the majority applicant, the response probability for the minority applicant was on average 11 percentage points lower. As can be seen from Table 1, the studies varied in the tested ethnicities, the year of study, and also the experimental design: Whereas in some studies only one application was sent per vacancy (between-design), in other studies various applications with different ethnicities were sent (within-design). Within-designs offer greater statistical power to detect discrimination (Charness, Gneezy, & Kuhn, 2012: p. 2). A downside is, however, the larger probability of being detected by suppliers as they receive at least two similar requests.

The pros and cons of these different experimental designs and also differences across ethnicities and time trends have already been discussed in the literature (e.g. Heckman, 1998; Vuolo, Uggen, & Lageson, 2016). We are, however, not aware of any studies focusing on the sampling techniques that are also summarized in Table 1. First, there is variation in the used sampling frame (sampling on the level of suppliers or apartments): Whereas in about half of the studies the researchers sent only one application to each supplier—although she or he might have advertised more apartments—(*supplier-*

[1] Observational data suffer from strong problems of unobserved heterogeneity: Ethnicity is correlated with many (unobserved) factors such as monetary resources, social networks and willingness to pay for housing (for an overview on different methods to measure discrimination see e.g. Pager & Shepherd, 2008).

sampling), in the other half of studies the researchers sampled every apartment with an equal probability (*apartment-sampling*).

Second, the length of the field period when apartments were sampled also differed across the studies, ranging from one week (Mazziotta, Zerr, & Rohmann, 2015) to more than half a year (Bosch, Carnero, & Farré, 2015). As we will demonstrate later on in more detail, when using very short sampling periods, in particular apartments with a high success rate have already disappeared from the market platform when the sampling period sets in. That is, new and short vacancies will be under-sampled, while apartments with a long advertisement time will be over-sampled. If the supplier-sampling or length bias have a meaningful impact on the observed level of discrimination, part of the cross-study variation shown in Table 1 would be artificial, simply caused by different sampling techniques.

Table 1. Sampling Techniques in E-Mail Correspondence Tests Published in 2010–2019

Study	Country	Nation wide [a]	Tested ethnicities	Design [b]	Sampling frame [c]	Field time (in months)	Cases (*N* Apart.)	Discr. rate (in ppts.) [d]
Acolin, Bostic, & Painter (2016)	France	y	African, Turkish, Polish, Portuguese /Spanish	b	s	2	1,800	13.9
Ahmed, Andersson, & Hammarstedt (2010)	Sweden	y	Arab	w	a	2	1,032	14.0
Andersson, Jakobsson, & Kotsadam (2012)	Norway	y	Arab	b	a	3	950	12.7
Auspurg, Hinz, & Schmid (2017)	Germany	n	Turkish	w	a	6	637	9.1
Baldini & Federici (2011)	Italy	n	Arab, Eastern European	b	a	4	3,676	15.0
Bengtsson, Iverman, & Hinnerich (2012)	Sweden	y	Arab	b	a	3	1,213	7.3
Bjornsson, Kopsch, & Zoega (2018)	Iceland	y	Polish	w	a	5	127	7.9
Bosch, Carnero, & Farré (2010)	Spain	n	Moroccan	b	s	3	1,809	12.7
Bosch, et al. (2015)	Spain	n	Moroccan	b	s	7	1,186	17.1
Bunel, Gorohouna, L'Horty, Petit, & Ris (2019)	France	n	Kanak	w	s	5	342	8.6
Carlsson & Eriksson (2014)	Sweden	y	Arab	b	a	5	5,827	10.7
Ewens, Tomlin, & Wang (2014)	USA	y	African American	b	s	2	14,237	9.3
Hanson, Hawley, & Taylor (2011)	USA	n	African American	w	s	4	4,728	6.3
Hanson & Santas (2014)	USA	n	Latino	w	s	1.25	3,072	-0.2

Heylen & Van den Broeck (2016)	Belgium	y	Moroccan, Turkish	b	s	2	1,769	18.6
Hogan & Berry (2011)	USA	y	African American, Asian, Arab, Jewish	w	a	4	1,124	2.9
Mazziotta, et al. (2015), Study 1	Germany	n	Turkish	b	s	0.25	336	29.3
Mazziotta, et al. (2015), Study 2	Germany	n	Turkish	b	s	0.25	456	14.0
Murchie & Pang (2018)	USA	y	African American, Arab, Latino	b	s	1.5	9,672	3.4
Oblom & Antfolk (2017)	Finland	n	Arab	w	a	5	800	13.7

Notes. [a] y: yes, n: no; [b] b: between, w: within; [c] sampling on the level of suppliers (s) or apartments (a); [d] risk difference between the response probability of the majority and the minority applicants in percentage points (ppts.).

3. Theoretical Background: Do Sampling Techniques Limit External Validity?

Researchers frequently make very general conclusions on the incidence of discrimination in the city, region, or even the whole country where they run their experiment. However, the generalizability of findings to a broader population depends on some requirements being met (for a general discussion of validity issues in experiments: Shadish, Cook, & Campbell, 2002). First, for valid descriptive results, the level of discrimination has to be the same in the studied and non-studied parts of markets. For this assumption to hold, characteristics that predict the level of discrimination have to be unrelated to the probability of being sampled (Bell & Stuart, 2016). Second, researchers are often also interested in causal relationships with characteristics of the applicants, suppliers or housing markets. For these associations to be externally valid, it would be necessary that one has included all important moderator variables (Shadish, et al., 2002).

The gold standard to ensure these assumptions would be to employ a (simple) random sample of tested units (Shadish, et al., 2002: ch. 3). However, there exists no official register of housing units that could be used for such purpose. Researchers have instead just relied on random or convenient samples of units that were advertised in newspapers or on internet platforms. In the following, we will first discuss how this might have led to a sampling bias. Second, we review theories on discrimination, with a special focus on the effects of characteristics likely affected by a sampling bias (size of supplier, advertisement time). In sum, this allows us to derive specific predictions on how sampling techniques affect the measurement of discrimination.

3.1. Do Sampling Techniques Lead to a Biased Sample of Housing Units?

As shown in Table 1, about half of the recent experiments used apartment-sampling, while the other half used supplier-sampling. Technically, the latter represents sampling *without replacement* on the level of

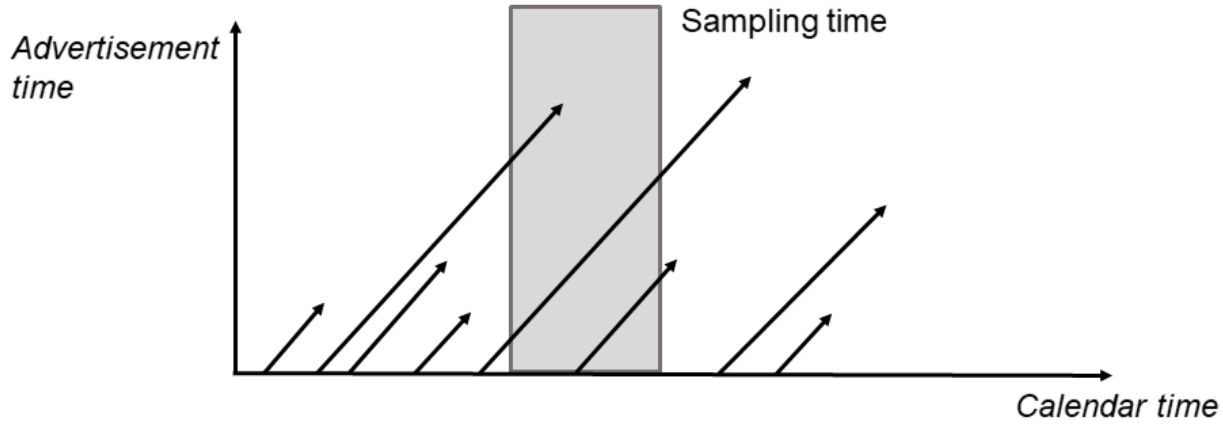
suppliers: Only the first drawn unit of each supplier is kept in the sample. As a consequence, offers by large, commercial suppliers are under-represented, whereas offers by small (often private) landlords are probably over-represented compared to the typical search process of a real apartment seeker.[2]

The second reason why the tested housing units might not be fully representative of the housing units real apartment seekers assess is the restricted field period. The tested housing units are necessarily sampled during a limited period, which typically lasts between one week and several months (see again Table 1). The shorter this time interval, the more likely it is that units that are advertised for a relatively long time are over-sampled. This length bias is graphically depicted in Figure 1. The x-axis of the *Lexis Diagram* shows the calendar time, while on the y-axis there is the time housing vacancies are advertised online. The vectors represent different offers. Field experiments use a cross-section, also called point sampling of these offers (see the shadowed area in Figure 1). Only units that are online during this restricted sampling period have a non-zero probability of being sampled. This means that the probability of being sampled is proportional to the time a unit is advertised online. Put differently, in particular units with a high survival time on the platform (long vacancies) are sampled. These units are those where it takes *per se* more time to find a tenant; or where landlords or agencies are especially picky in choosing their tenants. As can be seen from Figure 1, the shorter the sampling period, the more severe is this length bias.

One might argue that real apartment seekers also face this bias. They also look for housing only during a limited search interval and do not follow single apartments over time since they disappear from the market. However, as far as the length of the sampling window does not match the typical search interval of real apartment seekers, the sample would nevertheless be biased.

[2] A solution to restore a simple random sample would be the use of *design weights*, i.e. to weight all units by the inverse likelihood of being sampled, which is here the number of offers that were advertised by the same supplier. However, we are not aware of any studies where this technique was used.

Figure 1. Lexis Diagram: Association Between the Advertisement Time and Likelihood of Being Sampled



Notes. This diagram shows on the x-axis the calendar time, and on the y-axis the time housing offers are advertised online. The different vectors stand for different offers. The height of each vector displays their survival time online (i.e. advertisement time). As can easily be seen, cross-sections mainly catch offers with a long advertisement time (i.e. low success rate). This phenomenon is called *length bias*. Source: Authors based on van Es, et al. (2000: 307).

3.2. Does Discrimination Vary with the Size of Suppliers or the Length of Advertisements?

There are two main theories on discrimination: *First*, theories on *animus discrimination* or *tastes for discrimination* (e.g. Becker, 1957) assume that suppliers of apartments discriminate because contact with a disliked minority group, such as foreigners, would cause them a psychological disutility. To avoid this disutility, the suppliers are seen to bear economic losses, e.g. in the form of longer vacancies and forgone rental payments, until a tenant of the preferred majority group is found.[3] *Second*, theories on *statistical discrimination* (Aigner & Cain, 1977; Arrow, 1973, 1998; Phelps, 1972) assume that actors discriminate to lower problems of missing information. On the housing market, it is for instance difficult to predict to what extent possible tenants will provide stable rental payments. The ethnicity of apartment seekers could be used as a cheap proxy for this unknown characteristic, as migrants are known to have in general lower earnings and also a higher risk of being in unstable employment. However, in contrast to animus discrimination, these theories assume that discrimination helps suppliers *increase* their economic profits.

How are these mechanisms related to the size of suppliers and the time apartments are advertised online? Starting with the *size of suppliers*, there are plausible arguments in both directions that larger suppliers (that offer more housing units) are associated with a lower or higher level of discrimination. A prominent argument for a higher level of discrimination is that only the *big players* have enough market power (or assets) to survive in competitive markets when they engage in costly animus discrimination (e.g. Ashenfelter & Hannan, 1986). However, it might also be the other way round: Large

[3] A variant of this animus discrimination is *discrimination due to implicit attitudes*. Following these theories, discriminating actors act mainly unintentionally, based on quick (lexicographic) heuristics (Bertrand, Chugh, & Mullainathan, 2005; Quillian, 2006). In doing so, they follow unconsciously negative attitudes or (wrong) stereotypes on social groups, which are for instance revealed by psychological test instruments such as the Implicit Association Test (Greenwald, Nosek, & Banaji, 2003).

agencies probably only grew large because they did not engage in costly animus discrimination (Becker, 1957; Hellerstein, Neumark, & Troske, 1998). In addition, large suppliers are typically corporate agencies that act solely as brokers, and therefore have only limited contact with possible tenants (e.g. during the showing, the contract processing, and related correspondence). For this reason, tastes for discrimination certainly matter more for private landlords who start a longer-lasting relationship with their tenants. For similar reasons, corporate agencies can also be expected to have less impulse toward statistical discrimination, as it is not them but only apartment owners who would be affected by outstanding rental payments. Also the more formal and rule-driven processes within large agencies are expected to lower the risks of discrimination (Biddle & Hamermesh, 2013).[4] Indeed, there is already some evidence that real estate agents discriminate significantly less against minority applicants than private landlords (Flage, 2018). *We therefore expect that the typical over-sampling of small suppliers leads to an over-estimation of the overall extent of discrimination.*

Regarding the *time apartments are advertised online*, again different mechanisms come into play. First, suppliers without tastes can rent out their apartments more time-efficiently: They are *by definition* less picky, and hence should be able to find a tenant in a shorter time interval than landlords with tastes for discrimination. This means that apartments with long advertisement times are more likely those that are offered by discriminating actors. However, apartments with low success rates could also indicate a tight market. This could be a regional market with many vacancies, and at the same time only few apartment seekers; and/or market segments that have a relatively low demand (such as apartments that are relatively over-priced, offer few amenities and/or are located in poor neighborhoods). A standard prediction made by economic theories is that in tight markets mainly actors who act cost-effectively survive the competition, which are actors with no tastes for discrimination (Becker, 1957). A more realistic assumption is that landlords or agencies do not wait until they are driven out of the market but rather adapt their (discriminating) behavior to the market situation (Fernandez & Campero, 2014). Put differently, discriminating actors might not discriminate at any price, but only as far as the utility gained by discrimination outweighs the opportunity costs in the form of longer vacancies (Biddle & Hamermesh, 2013). If this is true, suppliers will show in general less willingness to discriminate, the more difficult it is to fill their dwelling, and hence the longer they have already advertised their apartment.

Therefore, one might expect two opposing effects related to the length bias. On the one hand, over-sampling units with low success rates could mean that one samples mainly the apartments of suppliers who discriminate. On the other hand, the lower the success rate, the less likely it is that actors still follow their tastes for discrimination. The net effect that results from these two countervailing mechanisms is

[4] This is also because the larger suppliers are more likely monitored and sanctioned for discrimination. For instance, legislation in Germany explicitly prohibits selecting tenants based on ethnicity only for larger suppliers that rent out apartments beyond their or their close relatives' place of living (see the German General Act on Equal Treatment [Law to implement the European Directive on the realization the principle of equal treatment], § 19 (http://www.ilo.org/dyn/natlex/natlex4.detail?p_lang=en&p_isn=77201).

difficult to predict. We therefore only *expect that the level of discrimination varies with the length of advertisement time*, and we will also explore the *possibility of non-linear patterns*.

4. Data

4.1. Sample of Housing Units

Our data were collected in a nationwide field experiment in 2015 in Germany, with the tested housing units being advertised on the most prominent online platform for housing units at that time.[5] On this platform, a large number of housing advertisements of both private as well as corporate suppliers were listed. The field experiment took place in two sampling periods, each of five days, in spring and winter 2015 (May 4th – May 8th, and November 30th – December 4th).

On each day in both sampling periods, a random sample of 500 advertised rental apartments was drawn, which resulted in a total sample size of 5,000 rental apartments. The sampling procedure was restricted to apartments with 2–4 rooms. Furthermore, each supplier was tested only once for each sampling period (i.e. we used a sampling without replacement on the level of suppliers). In doing so, we employed a typical sampling strategy for field experiments on the housing market (see again Table 1).

4.2. Experimental Design

The main treatment of the experiment was the ethnicity of the two male applicants (the gender of applicants was held constant to increase the statistical power regarding the effect of ethnicity). Each supplier got one inquiry from a Turkish and one inquiry from a German applicant (within-design). Both inquiries were sent in random order with a lag of about one hour between the two applications. Ethnicity was signaled by the names, included both in the signatures and the e-mail addresses of the applicants.[6] Turkish immigrants form the largest immigrant group in Germany (German Statistical Office, 2017) and were also tested in preceding correspondence studies (Auspurg, et al., 2017; Mazziotta, et al., 2015).

Besides their ethnic background, several other applicant characteristics were experimentally varied, such as (the extent of) information on the applicants' educational level (signaled by different occupations) and employment status.[7] The occupation was varied on three levels: There was either (1) no information; or the applicant indicated (2) an occupation that requires vocational training (indicated by six different occupations, e.g. the applicant mentioned that he is currently working as a nurse or electrician); or he mentioned (3) an occupation that requires a university degree (again signaled by six

[5] More information on the platform and used web scraping methods is available on request.

[6] For both German and Turkish applicants we selected 30 male names that were common and supposed to not signal a specific socioeconomic status, birth cohort, or invoke any other idiosyncratic associations. German names included, for example, Benjamin Buchholz, Maximilian Böhme, and Andreas Engelhardt. Turkish names were, for example, Volkan Sengül, Orhan Simsek, and Erol Tasdemir. All e-mail addresses followed the same format (name.lastname@provider), using a range of common providers, such as aol.de, gmail.com, and gmx.de.

[7] We also varied some further information, such as whether there is information on the household income and family status or not. For the analysis presented later we show exemplary results on the education and employment status. The main conclusions do not change when including further experimental treatments, but focusing only on these two main predictors for discrimination helps to keep the presentation and discussion of results more clear cut (results on the other experimental treatments are available on request).

different occupations, such as working as a medical doctor or architect). In case of statistical discrimination, one would expect that the gap in the response probabilities declines once information on an occupation requiring a high educational level is provided, as these occupations typically offer an especially high salary and job security (for a more detailed discussion and evidence see e.g. Auspurg, et al., 2017; Auspurg, et al., 2019b). The employment status was varied on four levels: (1) no information (to be again able to test for statistical discrimination); (2) permanently employed; (3) self-employed; or (4) working in the public sector. Being self-employed was thought to signal an insecure income. Being permanently employed or working in the public sector, in contrast, was thought to represent a particularly high level of income security due to the particularly high level of job security in the public sector in Germany.

The characteristics of the applicants were completely crossed with each other based on a *D*-efficient experimental design (for details: Auspurg & Hinz, 2015) that minimizes the correlations between the different treatment variables and at the same time maximizes their variance. Such a design allows for a maximum level of statistical power to estimate the impact of all treatments. Note that the ethnicity was always varied within the tested apartments, while other treatment variables could be the same or different for the two applications sent to one apartment. In order to minimize the risk of the experiment being detected, different wordings for the two inquiries were used, such as different salutations or orderings of the information displayed in the e-mails. The resulting pairs of e-mails (one always being of a Turkish and one of a German applicant) were randomly allocated to the sampled apartments. We carefully checked whether the randomization worked (i.e. approved that applicant characteristics were not correlated with any characteristics of the suppliers or regions where the experiments were done). We also assured that the different text versions did not evoke any idiosyncratic response patterns.[8] Figure 2 shows an example inquiry, with the experimentally varied applicant characteristics being underlined.

Figure 2. Sample Inquiry (Translated Version, Experimentally Varied Dimensions Are Underlined)

Dear Ms./Mr.,

I am very interested in the advertised apartment. My name is Volkan Sengül. I am permanently employed as an electrician, and I am looking for an apartment. I would be very grateful if you could offer me a showing.

Kind regards,

Volkan Sengül

Notes. In other e-mail variants, additionally information on the applicant having a partner or family and his income was provided. The whole list of variations and text phrases is available on request.

[8] The maximum correlation of a text version with the observed response pattern (i.e. discrimination) was $r = 0.025$, $p = 0.17$.

4.3. Market Data

We collected official statistics on the regions where the apartments were located. Besides that, we combined our experiment with market data captured on the internet platform itself. While the experiment took place in two sampling periods (in May 2015 and November/December 2015), each of five days, the platform was observed for a whole year, covering at least one month before and after the two sampling periods. Information on all advertised apartments was collected daily between March 2015 and February 2016, exporting each active advertisement with an automated web-scraping routine.[9] The outcome is a database that identifies the spells in which the apartments were advertised on the platform. All in all, we gathered spell data on 1,087,541 advertised apartments.

These data allow us to calculate the information we are interested in: the size of suppliers, measured by the number of offers they advertised during our one-year observation period; and the time offers were (already) advertised online (until we sent our e-mails). Suppliers that did not report any company name were coded as private landlords. For private landlords, we could not observe the exact number of offers belonging to the same supplier, because no distinct information was available on the platform.[10] Therefore, all private landlords were coded to have only one advertisement by design. Due to the high advertisement prices and the differential pricing offered by the online platform, this seems plausible: All suppliers could set up a corporate account that offered discount rates when having several offers, so landlords with multiple offers likely chose this option.

The advertisement time was measured in days. Short interruptions in advertisement times of up to 14 days were 'smoothed' (i.e. seen as one joint advertisement interval), as it is very unlikely that an apartment was rented and offered again within a time window of only two weeks. It is more plausible that it is still the same vacancy: Taking an apartment offline, organizing some showings, and putting it online again if there was no adequate tenant represents a cost-effective renting-strategy for landlords. Besides using a very long observation window, in some cases a left- or right-censoring of the advertisement time occurred (i.e. advertisements were already or still online when we started or finished monitoring the internet platform). We will provide results based on observations without censoring; robustness checks that also included censored advertisement times did not change any substantive conclusions (results are available on request).

We present results on apartments located in Western Germany. We decided to exclude Eastern Germany because the rental market as well as attitudes toward foreigners still strongly differ across both parts of Germany. To adequately capture these heterogeneities, one would have needed quite complex regression techniques (Auspurg, Brüderl, & Wöhler, 2019a). However, it is noteworthy that our substantive results still hold when including Eastern Germany. After deleting cases with missing

[9] Due to technical problems there were some gaps of two to nine days where we had no observations, spreading over the whole observation period (except for the experimental periods). This could have led to some censoring of advertisement times.

[10] Some private suppliers only indicated their first or second name, which does not allow for a valid identification of single suppliers.

observations on, for example, the advertisement time (mostly caused by left-censoring), 2,992 tested apartments (of formerly N = 3,932) remained in the sample. This number corresponds to 5,984 (=2,992 x 2) e-mail applications, half sent by a Turkish and half sent by a German applicant.

4.4. Identification Strategy

To see if there is the expected sampling bias in the size of suppliers and length of advertisements, we first contrast descriptive statistics on the sampled housing units (the *sample*) with statistics on the full population of housing units within the one-year observation window (the *market data*). To see if a possible sampling bias in these variables affects the measurement of discrimination or conclusions on theories on discrimination, we will then explore whether a) the measured level of discrimination and b) its association with applicants' characteristics (such as them providing more or less information) depends on the two apartment characteristics (size of supplier and advertisement time).

To *measure discrimination*, we follow standard procedures in the literature and look at the quantity of responses. Getting no response is definitely a rejection and in nearly all cases the response to an e-mail inquiry is a positive one (i.e. an invitation to a showing).[11] In our analyses, we will look at the apartment level and contrast the following three outcomes *j*:

(1) *Equal treatment (j = 0)*: Both the Turkish and German applicant get a response or both get no response.
(2) *Discrimination against the Turk (j = 1)*: Only the German (but not the Turk) gets a response.
(3) *Discrimination against the German (j = 2)*: Only the Turk (but not the German) gets a response.

The percentage of cases falling in category 2 (3) is the so-called gross discrimination rate against the Turkish (German) applicants. In addition to this, authors often report the net discrimination rate of minorities, which is defined as the difference between both discrimination rates (i.e. 'discrimination against the Turk' minus 'discrimination against the German;' for a detailed discussion of these discrimination measures see Wienk, Reid, Simonson, & Eggers, 1979a: p. 18). In our analyses, we will only explore the gross discrimination rates (i.e. the absolute outcomes 2 and 3). Looking at the gross discrimination rates allows for more fine-grained insights than when looking at the net discrimination rate that summarizes both outcomes (note that our results are, however, robust to the alternative net-measurement of discrimination). In addition, this analysis strategy mirrors typical analyses used in the literature (see e.g. Ross & Turner, 2005).

To see how both gross discrimination rates vary with the characteristics of interest, we first use non-parametric estimations (LOESS smoothers: Cleveland, 1979) that help to explore possible non-linear associations. In addition, we use multiple multinomial logistic regressions to contrast the three

[11] Results including qualitative information on the responses likely suffer from lower reliability, as the quality of responses is difficult to code (the subtle forms of discrimination are more difficult to uncover; see Hanson, et al., 2011). In our case, including qualitative information led to quite similar substantive conclusions (results on these robustness checks are available upon request).

different outcomes *j*: Equal treatment is used as the reference category (*j* = 0; for details on multinomial regression models see Greene, 2012: p. 763). Against this reference category of equal treatment, we estimate the likelihood of gross discrimination against the Turkish applicant (*j* = 1) and gross discrimination against the German applicant (*j* = 2).

Equation (1) shows this regression model. *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 different tested housing units ($i = 1, \ldots, N_{\text{apartments}}$). In all models, we include some control variables *C* that are known to influence the level of discrimination (percentage of foreigners in the county; apartment located in a city yes or no). *T* are the treatment variables in the form of the characteristics of the applicant besides their ethnicity, while *A* are the apartment characteristics of main interest, i.e. the size of the supplier as well as the time the apartment was advertised until treatment (i.e. until the e-mails were sent). Due to the skewed distribution of these variables, they enter the regressions in logarithmic specifications. Positive (negative) regression coefficients mean that the odds of the outcome *j* is increased (decreased) compared to the reference category. For our research aim the coefficient β_{Aj} in equation (1) is of most interest: A significant coefficient β_{Aj} would suggest that sampling bias in the tested housing characteristic *A* translates to a biased estimate in the level of discrimination (outcome *j*).[12] Positive (negative) effects mean that larger suppliers/longer advertisement times go along with higher (lower) odds of discrimination.

$$Logit(Y_i = j) = \beta_{0j} + \beta_{Aj}\ln A_i + \beta_{Tj}T_i + \beta_{Cj}C_i,$$
$$j = 0, 1, 2; \;\; i = 1, \ldots, N_{\text{apartm.}} \qquad (1)$$

In a second step (equation 2), we additionally include interaction terms between the treatment (applicants' characteristics) and apartment variables ($T \cdot \ln A$) in order to see if the sampling bias also affects the estimated effects of the treatment variables (and related tests of discrimination theories). A significant coefficient β_{TAj} of the interaction term would suggest that sampling bias translates to a biased assessment of the effects of treatment variables: Depending on the sample composition in regard to *A* (i.e. size of suppliers, advertisement times), different treatment effects would be estimated.

$$Logit(Y_i = j) = \beta_{0j} + \beta_{Aj}\ln A_i + \beta_{Tj}T_i + \beta_{TAj}T_i \cdot \ln A_i + \beta_{Cj}C_i,$$
$$j = 0, 1, 2; \;\; i = 1, \ldots, N_{\text{apartm.}} \qquad (2)$$

[12] Multinomial regressions estimate different regression coefficients for the different outcomes. In our case this allows to see whether the explaining factors for the discrimination against the Turk versus German (and possible measurement bias herein) differ.

The full regression models are shown in the Appendix (http://dx.doi.org/10.6084/m9.figshare.9890801). In the main text we will provide only visual representations of the effects of main interest. Multinomial regression coefficients give the effect on the logit and are difficult to interpret. We therefore present average marginal effects (AMEs), which indicate the mean increase in the probability of an outcome *j* (percentage points when multiplied by 100) that is caused by a marginal increase of the variable of interest, when averaged over all observations. All analyses are done with the statistical software Stata version 15 (StataCorp., 2015). For the local polynomial plots, the Stata procedure *lpolyci* was used, and the coefficient plots were created with the user-written Stata ado *coefplot* (Jann, 2014).

5. Results

5.1. Is there a Sampling Bias?

Table 2 shows descriptive statistics on the sample used in the experiment and on all offers advertised during our one-year observation window (see Section 4.3 for details). As expected, there were substantial differences. First, regarding the size and kind of suppliers the expected over-sampling of small, private suppliers occurred: In the sample, about half of all suppliers were private landlords (53%), while only 26% of all suppliers active on the internet platform were private. In particular the number of offers per supplier differed drastically between the sample (mean 22; median 1) and market data (mean 1,246; median 34). This huge difference was especially caused by the supplier-sampling: Even the largest supplier, although advertising 15,810 apartments in our one-year observation window, *by design* was sampled only once.

Table 2. Descriptive Statistics on the Sample and Market Data

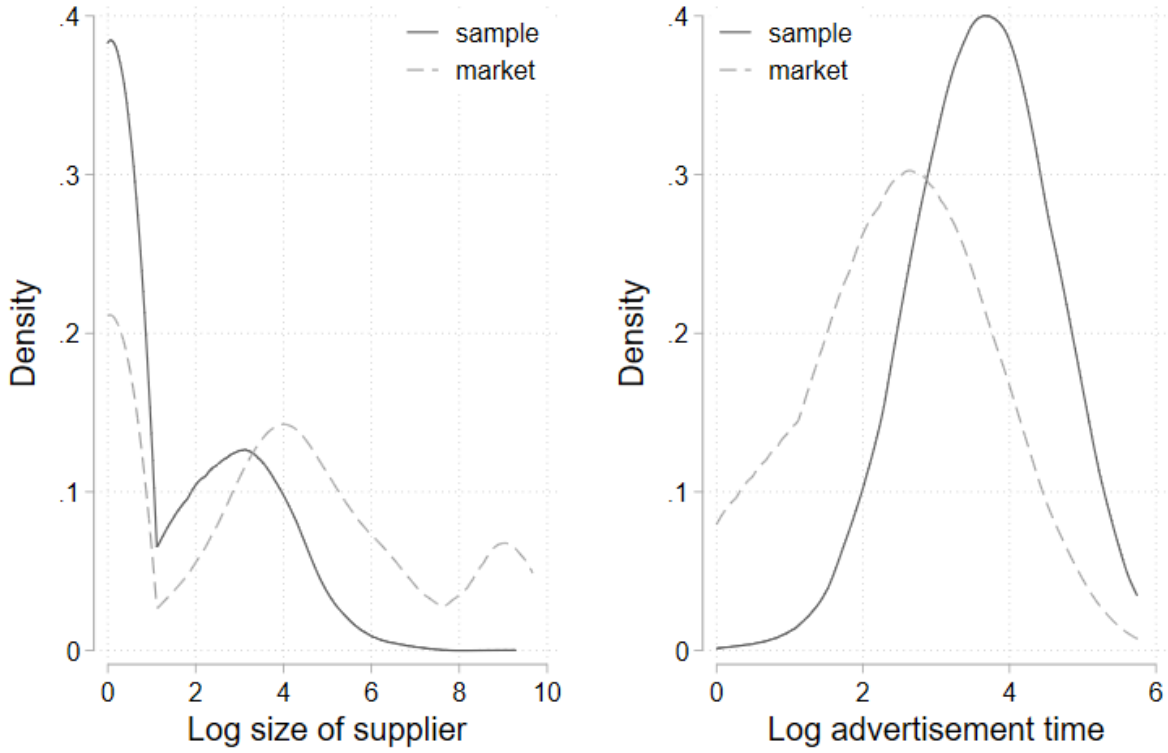
	(1) Sample Drawn for the Experiment		(2) Market Data: All Offers in the One-Year Observation Period	
	Mean	SD	Mean	SD
Supplier characteristics				
Private suppliers [a]	0.53		0.26	
Number of offers per supplier	22.35	203.10	1,246.13	3,303.80
Apartment characteristics				
Advertisement time (days) [b]	51.99	44.81	22.19	28.63
Size (sqm)	83.50	26.52	82.75	24.46
Prize (per sqm)	8.19	3.12	8.11	2.79
N	2,992		695,458	

Notes. [a] Coded as 'private' when there was no information on a company name (see Section 4.3). [b] Due to censored information on the advertisement time, we can calculate this statistic only based on 2,598 cases in our experimental data and 575,950 cases in the market data. In the analyses presented later on, we alternatively use the information on the advertisement time until our experimental treatment took place (i.e. we sent our e-mails), which allows for a slightly higher number of observations (*N* = 2,992 apartments).

Second, there was also strong evidence for the expected length bias. In the sample, the mean time an offer was advertised[13] on the platform was 52 days, which is more than twice the time found for the complete market data (22 days). In contrast, other characteristics of the apartments, like the cost of the rent or the size, differed only slightly between both data sources.

Because of the large skewness of the advertisement time as well as the number of advertisements offered by a supplier, both variables were log-transformed. In this logarithmic specification, large differences between sampled units and the whole population (market data) persist (see the kernel density estimates provided in Figure 3). Furthermore, it can be seen that especially the largest suppliers in the right tail of the market data were under-represented in the experimental data in favor of very small suppliers (with mainly only one offer; see the kernel density estimate in the left panel). For the advertisement time, the distribution in the experimental data is shifted to the right, showing an over-representation of advertisement times of at least 20 days (re-transforming the logarithmic value 3 gives $e^3 = 20$).

Figure 3. Kernel Density Estimate of the Size of Supplier (Left Panel) and Advertisement Time (Right Panel) by Data Source (Sample vs. Market Data)



Notes. Size of supplier: number of offers per supplier. Advertisement time: overall time (in days) the apartment was advertised on the internet platform. Both graphs were produced using the Stata command *kdensity*. For the number of cases see Table 2.

[13] This is the overall time an advertisement was online. In the models presented later on we will use the advertisement time until experimental treatment (i.e. until we sent our e-mails).

To see if the observed sampling bias does not only hold for our unique sample of housing units, we also run simulations of different samples (see Table A1 in the Appendix). Simulating both sampling frames (supplier versus apartment level) in combination with different sampling periods (ranging from one week to six months), we always found the expected biases: Sampling on the supplier level led to an under-representation of large suppliers, while with longer sampling periods the observed advertisement time decreased. This makes us confident that *there is indeed a strong sampling bias in terms of over-representing small suppliers and offers with long advertisement times*.

5.2. Does Sampling Bias Affect the Level of Discrimination?

Table 3 shows a cross-tabulation of responses to the Turks' and Germans' applications. One can see that in most of the tested apartments (81.8%=32.6% + 49.2%) the suppliers treated the Turkish and the German applicant equally: In 32.6% of the cases they answered neither of the two e-mails; and in 49.2% of the cases they responded to both e-mails. These cases will form the reference category in our multinomial regression analyses (see Section 4.4). In the remaining cases, unequal treatment or *discrimination* occurred. The observed *gross discrimination rate against the Turks* is 14.4% (in 14.4% of all cases only the German applicant got a response). The *gross discrimination rate against the German* is 3.8% and hence quite lower (only in 3.8% of all cases the Turkish applicant alone got a response). Taken together, this means that there was a net discrimination against Turks of 14.4 - 3.8 = 10.6 percentage points. This discrimination rate is close to that reported in other field experiments on Germany, and it also comes close to the risk difference in majority and minority response probabilities reported in our literature review in Section 2.[14]

Table 3. Response Patterns and Resulting Discrimination Rates

		German applicant (G)		Total
		No response	Response	
Turkish applicant (T)	No response	975 (32.6%)	430 (14.4%)	1,405 (47.0%)
	Response	114 (3.8%)	1,473 (49.2%)	1,587 (53.0%)
Total		1,089 (36.4%)	1,903 (63.6%)	2,992 (100.0%)

Notes. This table shows the obtained response patterns for the 2,992 apartments tested in Western Germany with full information on the supplier size and advertisement time until treatment (i.e. until we sent our e-mails). The

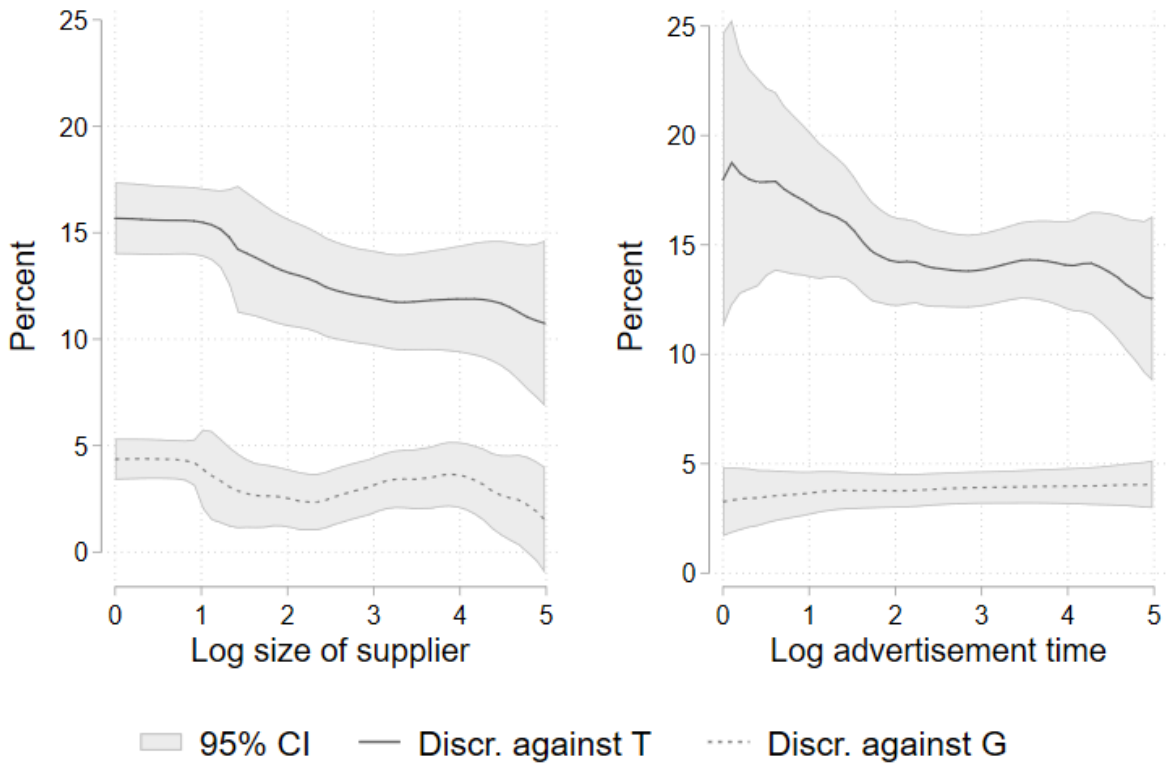
[14] These risk differences are recommended to be used in meta-analyses, as reported in Section 2. For the research aims in this paper, however, it seems more reasonable to rely on the simpler (and easier to interpret) gross discrimination rates.

14.4% (3.8%) of cases with only a response to the German (Turkish) applicant define the gross discrimination rate against Turks (Germans).

Do the discrimination rates vary by the size of supplier and/or length of advertisement times? Before switching to regression analyses, we present in Figure 4 non-parametrical estimates that rely on fewer assumptions and help to identify possible non-linear patterns. Due to the very skewed distribution of both independent variables (see Table 2) we use logarithmic specifications, and restrict the analyses to values with sufficient observations for stable estimates. The shadowed areas display the 95% confidence intervals. In the left panel one can see that with an increasing size of suppliers the gross discrimination against Turks declines: From 16.0% for suppliers that offer only one apartment (which is about half of our sample) to 12.5% for the largest suppliers (that offer up to ln(5), which is about 150 offers). Given a mean discrimination rate of 14.4%, this decline by more than four percentage points might be seen as substantial in effect size; it is, for example, of a similar level as differences found across countries (see Table 1 and Auspurg, et al., 2019b). Unequal treatment against the German applicant in contrast remained constant at 4% over the whole range of supplier sizes. Therefore, the net discrimination rate, indicated by the distance between both discrimination rates, also declined with the increasing size of suppliers.

For the advertisement time (time online until experimental treatment, i.e. until we sent our e-mails), a similar pattern was observed (see the right panel). Unequal treatment against the Turkish applicant decreased from around 17% for very short advertisement times (1 day) to about 13% for very long advertisement times (of about 150 days). In contrast, unequal treatment against the German applicant was always around 4%. Hence, again, there were marked differences in the observed level of discrimination. However, due to the very large confidence intervals, the variations in the level of discrimination probably do not reach statistical significance.

Figure 4. Local Polynomial Smooth Curves of the Discrimination Rates Against the Turkish (T) and German (G) Applicant Over the Size of Supplier (Left Panel) and Advertisement Time (Right Panel)



Notes. The gray areas show the 95% confidence intervals. G: German; T: Turkish. The left panel shows the log-number of offers per supplier on the x-axis, the right panel the log-number of days an apartment was already advertised online until treatment. Both figures are based on N = 2,992 apartments. The figures were produced using the Stata command *lpolyci*.

To back up these results, we estimated parametric multinomial logistic regressions. The full models with all logit coefficients and information on the model fit can be found in the Appendix (Table A2, which shows regressions with a stepwise inclusion of covariates). Figure 5 only displays the effects of the apartment characteristics of main interest, the size of the supplier and advertisement time, while controlling for applicant characteristics besided ethnicity and some further control variables (see the Figure notes). Effects are reported as AMEs. Looking first at the 'discrimination against the Turkish applicant' (left panel), one can see that (compared to the reference of 'equal treatment') the likelihood of this outcome lessens with the size of suppliers (by on average 0.92 percentage points if the supplier size is increased by one log-unit: AME = -0.0092; $p < 0.01$; see Table A2 and A3 in the Appendix for the exact estimates).[15] This mirrors the descriptive results reported so far.[16] When switching from the

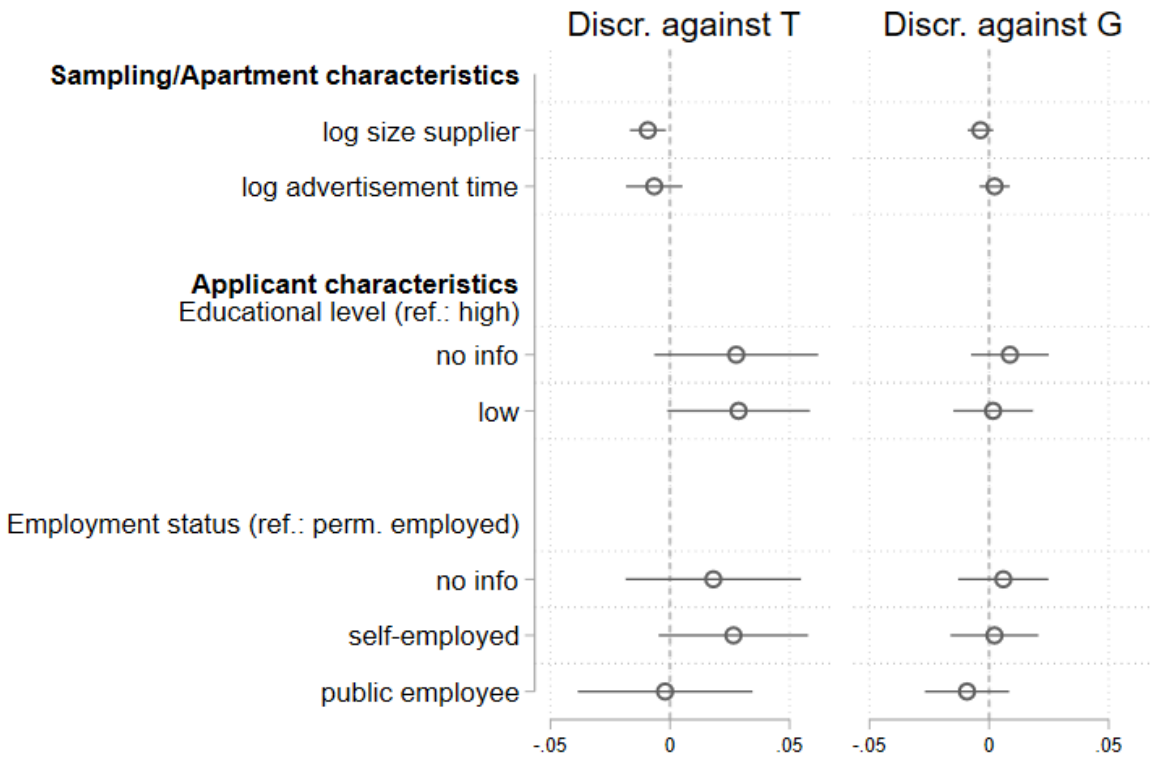
[15] Table A2 shows the logit estimates that were chosen for reporting the models with interaction effects, as interaction effect estimates are misleading when linearizing nonliear models by AMEs (Norton, Wang, & Ai, 2004). Table A3 shows in addition the AMEs presented in the Figures.

[16] When controlling for 'private landlord' the effect was no longer statistically significant, meaning that the effect was mainly driven by the higher discrimination related to small private landlords. In the analyses shown here, we do not control for private versus commercial landlords to prevent an 'overcontrol bias' caused by the inclusion of mediator variables (Elwert & Winship, 2014).

smallest (only one offer) to the largest suppliers (~150 offers), the discrimination rate is predicted to decline on average by 4.61 percentage points (as -0.0092 · ln(150) · 100 = -4.61).

The time an offer was advertised online (until treatment) also showed a negative effect on discrimination against the Turkish applicant, but this effect was not statistically significant (the confidence intervals overlap with the zero-line). For discrimination against German applicants, no substantial effects are found. All these results were robust to other linear and non-linear specifications (the reported log-specification of both variables provided the best model fit). *All in all, we can conclude that over-sampling small suppliers (which is mainly caused by a supplier-sampling) leads to a small, significant over-estimation of discrimination against minority applicants; while short field periods (i.e. over-sampling long advertisements) tends to lessen the observed discrimination rates, but only to a non-significant degree.*

Figure 5. Multinomial Logistic Regressions, AMEs of Predictors with 95% Confidence Intervals for the Outcomes 'Discrimination Against the Turkish Applicant' and 'Discrimination Against the German Applicant' (Reference: 'Equal Treatment')



Notes. Displayed are AMEs. For the outcome 'Discrimination against the Turkish applicant' ('Discrimination against the German applicant') effects of the characteristics of the Turkish (German) applicant are shown. Characteristics of the other applicant are controlled, as well as the percentage of foreigners in the area and whether the apartment was located in a city yes or no. See Table A2 in the Appendix for logit estimates for all variables and Table A3 in the Appendix for the AMEs displayed in this Figure. Results are based on *N* = 2,992 apartments.

Figure 5 also reports the effects of some applicants' characteristics. For the Turkish applicants (see again the left panel), one can see that both providing no information or indicating a low educational level (signaled by the occupations mentioned in the e-mails) tended to increase the probability of being discriminated against. The latter effect is significant at the 10% level ($p = 0.065$). Regarding employment status, no information also tended to increase the risk of discrimination against Turks, and there was also a tendency that self-employed Turks were more likely discriminated against compared to the reference of permanently employed applicants (this effect was again significant at the 10% level: $p = 0.075$). In regard to discrimination against the German applicant, again no substantial effect was found. In sum, this means that there was only weak evidence for statistical discrimination against Turks: Information that signals a higher or more stable income tended to lower the risk of discrimination; but overall the effects were not strong enough to reach statistical significance.

5.3. Does Sampling Bias Affect the Effects of Other Treatment Variables?

In a final step, we are interested in whether sampling bias affects the measurement of the effects of applicants' characteristics (treatment variables besides ethnicity). To find out, we have to see whether the effects of these variables are moderated by the size of suppliers or advertisement times.

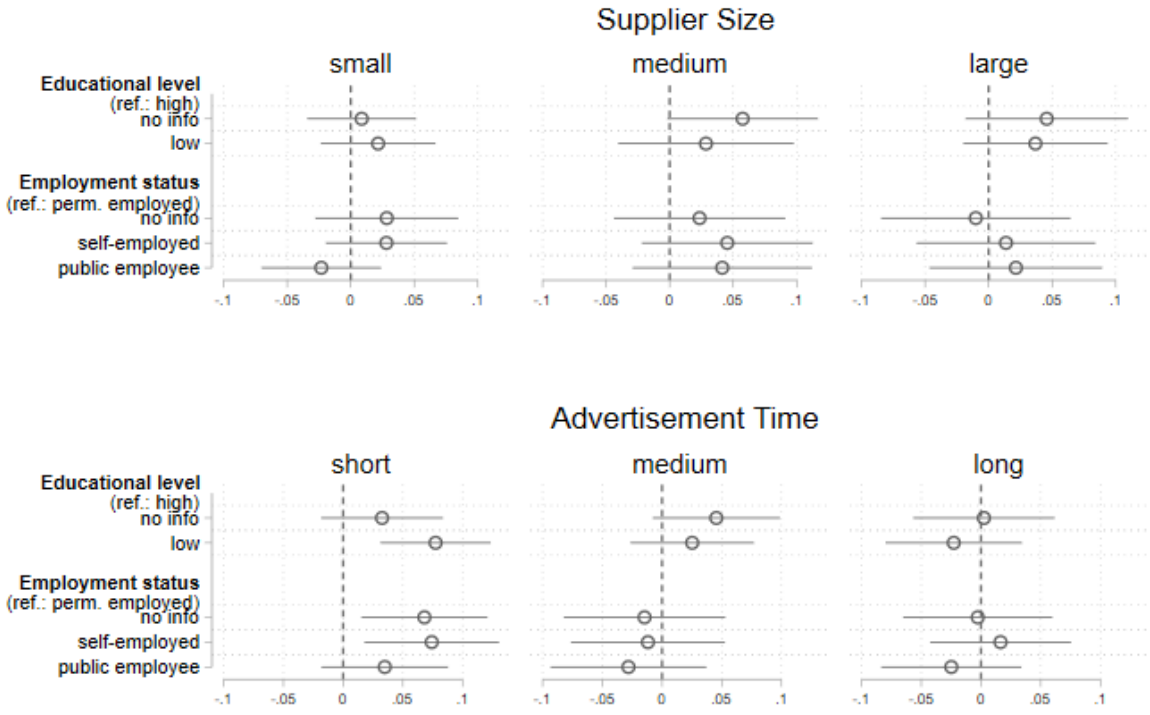
To ease interpretation, and also to be able to observe possible non-linear patterns, we present our results as split sample analyses by three strata of the size of supplier and length of advertisements. For the size of suppliers, we contrast the 56% of suppliers with only one offer with two other strata of about equal size (each containing about 22% of suppliers: 2–17 offers; and 18–10,684 offers); for the length of advertisements, we contrast the three terciles (1–12 days; 13–29 days; 30–245 days). In the Appendix, we provide in addition pooled regression analyses that include the interaction terms of treatment and apartment variables ($T \cdot \ln A$), which allow testing of whether moderation effects by apartment characteristics are statistically significant (Table A2). In the following, we only show results on the discrimination against Turks, where we found so far the strongest effects for treatment variables; results on discrimination against Germans are provided in the Appendix (Figure A1 and Table A5).

Figure 6 shows the results. For the size of the supplier (top panel), no clear pattern emerges: The effects of educational levels seem to be somewhat stronger for larger suppliers, but for employment status it is the opposite (i.e. smaller effects for larger suppliers). All these differences across samples are probably only due to random variations. This is supported by the non-significant interaction of both treatment variables with the size of suppliers in pooled regressions (see Table A2 in the Appendix).

Regarding the split samples by advertisement time (bottom panel), it is noteworthy that there are for nearly all treatment variables the strongest effects in the stratum with very short advertisement times (up to 12 days). In this sample based on relatively short (and hence new) vacancies, the effects of several treatment variables reach statistical significance. One can, for example, observe that for a low (compared to high) educational level the risk of being discriminated against is significantly higher (by 7.7 percentage points, AME = 0.077). With the other two samples one would, however, conclude that educational background does not make a significant difference. Similar patterns emerge for employment

status. These observations suggest that very short field periods (that mainly catch the long advertisement times) somewhat underestimate the incidence of statistical discrimination (i.e. discrimination depending on the amount and kind of information on the applicants). The observation is also in line with our assumption that landlords become less picky when it takes more time to fill their apartments. However, pooled estimates with interaction effects indicate that the differences across strata are not statistically significant (see again Table A2 in the Appendix). In a practical sense, the results nevertheless suggest that different samples can lead to quite different conclusions—although some of the differences are probably simply caused by sampling error (we return to this in our discussion). Note that for the discrimination against Germans, no remarkable patterns or statistically significant differences were found (see Figure A1 and Table A5 in the Appendix). *Overall, we conclude that sampling techniques and in particular the length of field periods have some impact on the estimated treatment effects; however, in our case study the moderation of treatment effects by size of supplier or advertisement time is too small to reach a statistically significant level.*

Figure 6. Multinomial Logistic Regressions, AMEs for the Outcome 'Discrimination Against the Turkish Applicant' with 95% Confidence Intervals for Split Samples by Supplier Size and Advertisement Time



Notes. Displayed are AMEs estimated by separate regressions for small (only one offer), medium (2–17 offers) and large suppliers (at least 18 offers); (top panel); and short (1–12 days), medium (13–29 days) and long advertisement times (at least 30 days) until treatment (bottom panel). Additionally controlled for: applicant's family status (with partner, with family, vs. single), and all characteristics of the applicant of the other ethnicity who e-mailed the same supplier, as well as the percentage of foreigners in the area and whether the apartment was located in a city yes or no. See Table A4 in the Appendix for the exact estimates. All estimates are based on at least $N = 658$ observations.

5.4. Robustness Checks

We presented analyses on the apartment level, where we estimated multiple multinomial regressions to predict both the likelihood of discrimination against Turks and Germans against the reference category of equal treatment (i.e. both getting a response or both getting no response). Substantive results are robust to using an alternative reference category (both getting no response; while treating 'both getting a response' as a separate outcome).

An alternative approach used in discrimination research is to use the e-mail inquiries as units of analysis and to estimate regressions with a dichotomous dependent variable (is there a response, yes or no). In this approach, the occurrence of discrimination is observed by a significant main effect of Turkish ethnicity. As pairs of e-mails are nested within the same apartment, multilevel regressions are appropriate (random effects regressions). When using this alternative approach (where we tested effects on the level of discrimination by bivariate interactions of Turkish ethnicity with supplier size/advertisement time, and the moderator hypotheses by three-way interactions of the Turkish ethnicity, applicants' characteristics, and supplier size/advertisement time), the main conclusions were very similar: There were again variations in the level of discrimination with a maximum of four percentage points difference in the levels (again the level of discrimination was found to be higher for small suppliers and short advertisement times), and treatment effects differed marginally by these moderator variables. Besides similar effect sizes to the ones reported in our results section, none of these variations reached statistical significance at the 5% level. This is probably due to the much lower statistical power when testing interaction effects instead of main effects (Cohen, 1988: pp. 367–369). Nevertheless, these robustness checks found the same direction and effect sizes.

To ensure that we did not overlook non-linear effects, we not only employed non-parametric analyses (LOESS curves, as presented in Figure 4) but also tested different parametric specifications, such as regressions with polynomial terms and categorical variables for the sampling/apartment characteristics.

Finally, we also tested whether weighting regression analyses by the size of supplier in order to correct for the different sampling probabilities leads to different conclusions. Again, this was not the case: Weighting decreased the impact of our market variables on discrimination to a small extent but did not change substantive conclusions on the effects of treatment variables.

6. Summary

The aim of our study was to analyze how sampling strategies affect the external validity of field experiments. Combining a large-scale field experiment on ethnic discrimination in the German rental housing market with rich data gathered from the platform where the apartments were advertised, we were able to test how both sampling on the level of suppliers (instead of apartments) and point sampling (i.e. sampling a cross-section of housing units advertised during a limited field period) affects the observed level and mechanisms underlying ethnic discrimination. Although these sampling techniques

are frequently used in field experiments, we are not aware of any other study that systematically discussed or analyzed these issues. We summarize our main findings in three points.

First, when comparing our sample to the larger population of housing units where our sample was drawn from, there was clear evidence of a sampling bias: As expected, sampling on the level of suppliers led to a huge over-representation of small suppliers with only few offers. At the same time, there was a strong over-representation of apartments with relatively long advertisement times. These aspects suggest that standard field experiments over-sample apartments that are offered by private landlords or other agencies with a small market power, and/or offers that stay a relatively long time on the market.

Second, we analyzed how both issues affect the observed level of ethnic discrimination. Empirically, the over-sampling of small/private suppliers particularly tended to increase the amount of discrimination observed in our study. These differences in effect sizes were, however, moderate: Even when contrasting the most extreme strata in our sample (containing only the apartments that were advertised by the smallest or largest suppliers), the found discrimination rate against Turks was within a relatively narrow range of 12.5% to 16.0%. Similarly, long advertisement times only slightly tended to go along with lower discrimination rates (differences in effect sizes were not statistically significant).

Third, we also analyzed whether the sampling bias might have affected the observed risk factors for discrimination. We tested characteristics of applicants (varying information on their educational and employment status) that were commonly used to explore statistical discrimination. We found some moderation effects, especially for the length of advertisements: Effect sizes tended to be a little bit stronger for offers that were advertised only for a relatively short time when our experiment took place, and only for these (new) offers we oberserved effects that point to statistical discrimination. However, none of the moderation effects were found to be statistically significant. All in all, the observed levels and incidence of (statistical) discrimination seemed to be mostly immune to deviations from 'representative', random samples of housing units.

7. Conclusions

Putting all these findings together, we conclude that the results of field experiments seem to be remarkably robust against sampling bias in terms of supplier size and advertisement time. Not surprisingly, descriptive findings were found to be more strongly affected than multivariate results: According to our estimates, the level of discrimination might be up to four percentage points larger or smaller, when one studies extreme samples that over-represent specific suppliers or advertisements. Given a baseline discrimination rate (observed on average in our sample) of about 14%, this variation might be classified as substantial. When comparing studies over time or done in different countries, sampling bias might be large enough to obscure time trends or cross-country differences that were so far found to be of similar size.

We nevertheless conclude that in particular the studied treatment effects in terms of applicants' characteristics (e.g. information on their employment status) were remarkably robust. Given that these

variables are of main interest to advance our knowledge on underlying mechanisms (animus or statistical discrimination) and also to advance interventions (Neumark, 2012), this is good news. Given the very similar general patterns found across all strata, and given the large number of variables tested in our study, most differences likely occurred just by chance (i.e. resulted from random sampling error). This rather points to the necessity of consolidating findings with replications based on other (similar or different) subsamples than to strong biases caused by sampling techniques.

Nevertheless, to increase the comparability of findings across experiments, one should try to be mostly transparent on the sampling techniques used. Only this would allow to include this information in future meta-analyses (the existing ones controlled only for few other design variables: Auspurg, et al., 2019b; Flage, 2018). A detailed documentation of the sampling procedure would also allow the use of design weights (for the application in surveys see Lavallée & Beaumont, 2015). Re-weighting the sample or using samples on the level of apartments from the beginning on certainly more closely matches the search strategies used by real apartment seekers.

In this context, it has to be stressed that the generalizability of research findings to 'real world processes' does not necessarily require a random sample (Salganik, 2017). Often it is difficult or even impossible to define a clear population and hence sampling frame; or doing so could result in excessive costs (Brewer & Crano, 2014). The results of experiments only do not generalize to other settings when there are differences that moderate the association between treatment and outcome variables. To ensure a high external validity of findings, it is in general important to study not only a broad range of different units, but also treatment variables, outcomes and experimental settings (Shadish, et al., 2002). Combining field experiments with market data is herein not only beneficial in terms of safeguarding the external validity of findings, but also in identifying important moderator variables: Observing variance across (market) contexts can provide important insights for advancing theories (Brewer & Crano, 2014). Market data might also simply be sampled from the platforms (without using larger observation windows); for example, one might use the indicated names of companies to collect further information on the suppliers.[17] Nevertheless, if the goal of research is to describe the 'true' level of discrimination individuals face, one might try to match their search processes as closely as possible to ensure that one really has mirrored the sample composition of all possible moderator variables.

In the last section, we want to discuss how our study points to directions for future research. *First*, and probably most important, we only provided evidence on one single case study in Germany. We tried to allow for more general conclusions by means of simulations to ensure that our observed sampling bias is not only bound to the one, idiosyncratic sample drawn for our field experiment. However, the results nevertheless might not generalize to other countries. In a meta-analysis on the

[17] When doing so, one has to consider ethical concerns. In addition, one should carefully check the validity of those 'big data'. In particular information provided by the platform itself should be regarded with caution. For example, especially commercial suppliers might pay for having their offers ranked as 'new' even though they have advertised them already for quite some time. Also, 'time drifts' in the way information is provided or idiosyncrasies due to the used algorithms ('algorithmic confounding') represent additional sources of errors for these data (Salganik, 2017).

housing market, Germany stands out as having a slightly higher amount of discrimination than is observed in other countries (Auspurg, et al., 2019b). This suggests that there are cross-country differences that moderate discrimination, and these differences might also moderate the associations analyzed in our case study. We therefore encourage researchers to repeat similar studies in other countries.

Second, our study was restricted to one online housing market. This might pose an even severe threat to the external validity of findings, as real apartment seekers likely use additional sources of information on vacancies, such as newspapers or social networks. That said, we are not aware of any study that compared samples based on different (offline and online) media. The bias that is caused by focusing only on one specific search strategy or (online) platform is probably more problematic than the bias we analyzed in this study. For instance, it seems plausible that in newspapers and in particular in social networks more units made available by small private (or older) landlords are advertised, and that in particular in social networks tastes for discrimination prevail (for some qualitative evidence on the often very subtle processes of discrimination that take place in personal interactions, see e.g. Krysan & Crowder, 2017). Without systematic research on the effects of sampling frames, one can only speculate on the direction and size of a possible bias. Another fruitful direction for future research would therefore be to study how discrimination (and underlying mechanisms) vary with the used (social) media to identify available housing units.

Third, although we collected 'big data' on the internet platform with an observation method that was 'always on' during our one-year observation period (Salganik, 2017), and hence gathered very fine-grained spell data with a longitudinal dimension, the field experimental data represent only a cross-section: Each vacancy was tested only at one specific point in time. This hampers the identification of different mechanisms that might be bound to the length of advertisements. The observed reduction of discrimination over the course an apartment was unsuccessfully advertised online could indicate that discriminating actors gave in to market pressure, and started to make compromises on the preferred attributes of tenants. However, to truly test this mechanism one would need longitudinal data also on the experimental side, i.e. testing single apartments several times over the course they are advertised online.[18]

Finally, combining field and market data can certainly provide promising new insights into the conditions and mechanisms underlying discrimination, or also other social interactions that are embedded in market structures. In this article we focused on methodological issues, but one could use similar combinations of field experiments with market data also to study substantive questions of interest in (discrimination) research, such as how the tightness of markets or market power of different large suppliers affect their (discriminating) actions (for exemplary discussion on these issues see e.g.

[18] However, this would also bring along some ethical issues. Testing the same supplier multiple times puts a higher, possibly disproportionately high, burden on this supplier. On the other hand, such an approach could provide unique insights into the mechanisms underlying discrimination. Ethics committees might decide what is reasonable.

Ashenfelter & Hannan, 1986; Baert, Cockx, Gheyle, & Vandamme, 2013; Carlsson, Fumarco, & Rooth, 2018). We hope that our case study was also stimulating in that way.

Acknowledgments

For helpful suggestions, we thank participants of the conference "Analytical Sociology" at Venice International University in 2017. We are grateful for comments on earlier versions we received from two anonymous reviewers and from the editors. Maximilian Sonnauer helped us in compiling the database for the field experiments.

Data Note

We used data collected in the research project "Ethnic Discrimination and Segregation in German Housing Markets" funded by a small non-profit foundation in Germany, the Wolfgang and Anita Bürkle foundation. Replication files (Stata do-files and the field experimental data) can be found in the Supplemental Material accompanying this article.

References

Acolin, A., Bostic, R., & Painter, G. (2016). A Field Study of Rental Market Discrimination Across Origins in France. *Journal of Urban Economics, 95*, 49-63.

Ahmed, A. M., Andersson, L., & Hammarstedt, M. (2010). Can Discrimination in the Housing Market Be Reduced by Increasing the Information about the Applicants? *Land Economics, 86*, 79-90.

Aigner, D. J., & Cain, G. G. (1977). Statistical Theories of Discrimination in Labor Markets. *Industrial and Labor Relations Review, 30*, 175-187.

Andersson, L. R., Jakobsson, N., & Kotsadam, A. (2012). A Field Experiment of Discrimination in the Norwegian Housing Market: Gender, Class, and Ethnicity. *Land Economics, 88*, 233-240.

Arrow, K. J. (1973). The Theory of Discrimination. In O. Ashenfelter & A. Rees (Eds.), *Discrimination in Labor Markets* (pp. 3-33). Princeton: University Press.

Arrow, K. J. (1998). What Has Economics to Say About Racial Discrimination? *Journal of Economic Perspectives, 12*, 91-100.

Ashenfelter, O., & Hannan, T. (1986). Sex Discrimination and Product Market Competition: The Case of the Banking Industry. *The Quarterly Journal of Economics, 101*, 149-173.

Auspurg, K., Brüderl, J., & Wöhler, T. (2019a). Does Immigration Reduce the Support for Welfare Spending? A Cautionary Tale on Spatial Panel Data Analysis. *American Sociological Review, 84*, 754-763.

Auspurg, K., & Hinz, T. (2015). *Factorial Survey Experiments*. Thousand Oaks, California: Sage.

Auspurg, K., Hinz, T., & Schmid, L. (2017). Contexts and Conditions of Ethnic Discrimination: Evidence from a Field Experiment in a German Housing Market. *Journal of Housing Economics, 35*, 26-36.

Auspurg, K., Schneck, A., & Hinz, T. (2019b). Closed Doors Everywhere? a Meta-Analysis of Field Experiments on Ethnic Discrimination in Rental Housing Markets. *Journal of Ethnic and Migration Studies, 45*, 95-114.

Baert, S., Cockx, B., Gheyle, N., & Vandamme, C. (2013). Do Employers Discriminate Less If Vacancies Are Difficult to Fill? Evidence from a Field Experiment. *Institute for the Study of Labor, IZA Discussion Paper No. 7145*.

Baldini, M., & Federici, M. (2011). Ethnic Disrimination in the Italian Rental Housing Market. *Journal of Housing Economics, 20*, 1-14.

Ball, M. (2016). Housing Provision in 21st Century Europe. *Habitat International, 54*, 182-188.

Becker, G. S. (1957). *The Economics of Discrimination* (2 ed.). Chicago: University of Chicago Press.

Bell, S. H., & Stuart, E. A. (2016). On the "Where" of Social Experiments: The Nature and Extent of the Generalizability Problem. *New Directions for Evaluation, 2016*, 47-59.

Bengtsson, R., Iverman, E., & Hinnerich, B. T. (2012). Gender and Ethnic Discrimination in the Rental Housing Market. *Applied Economic Letters, 19*, 1-5.

Bertrand, M., Chugh, D., & Mullainathan, S. (2005). Implicit Discrimination. *American Economic Review, 95*, 94-98.

Bertrand, M., & Duflo, E. (2017). Field Experiments on Discrimination. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of Economic Field Experiments* (pp. 309-393). Amsterdam: North-Holland.

Biddle, J. E., & Hamermesh, D. S. (2013). Wage Discrimination over the Business Cycle. *IZA Journal of Labor Policy, 2*, 1-19.

Bjornsson, D. F., Kopsch, F., & Zoega, G. (2018). Discrimination in the Housing Market as an Impediment to European Labour Force Integration: the Case of Iceland. *Journal of International Migration and Integration, 19*, 829-847.

Bosch, M., Carnero, M. A., & Farré, L. (2010). Information and Discrimination in the Rental Housing Market: Evidence from a Field Experiment. *Regional Science and Urban Economics, 40*, 11-19.

Bosch, M., Carnero, M. A., & Farré, L. (2015). Rental Housing Discrimination and the Persistence of Ethnic Enclaves. *SERIEs, 6*, 129-152.

Brewer, M. B., & Crano, W. D. (2014). Research Design and Issues of Validity. In C. M. Judd & H. T. Reis (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (2 ed., pp. 11-26). Cambridge: Cambridge University Press.

Bunel, M., Gorohouna, S., L'Horty, Y., Petit, P., & Ris, C. (2019). Ethnic Discrimination in the Rental Housing Market: An Experiment in New Caledonia. *International Regional Science Review, 42*, 65-97.

Carlsson, M., & Eriksson, S. (2014). Discrimination in the Rental Market for Apartments. *Journal of Housing Economics, 23*, 41-54.

Carlsson, M., Fumarco, L., & Rooth, D.-O. (2018). Does Labor Market Tightness Affect Ethnic Discrimination in Hiring? *Institute for the Study of Labor, IZA Discussion Paper No. 11285*.

Charness, G., Gneezy, U., & Kuhn, M. (2012). Experimental Methods: Between-Subject and Within-Subject Design. *Journal of Economic Behavior & Organization, 81*, 1-8.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association, 74*, 829-836.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.

Drever, A. I., & Clark, W. A. V. (2002). Gaining Access to Housing in Germany: The Foreign-minority Experience. *Urban Studies, 39*, 2439-2453.

Elwert, F., & Winship, C. (2014). Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual review of sociology, 40*, 31-53.

Ewens, M., Tomlin, B., & Wang, L. C. (2014). Statistical Discrimination or Prejudice? A Large Sample Field Experiment. *Review of Economics and Statistics, 96*, 119-134.

Fernandez, R. M., & Campero, S. (2014). Does Competition Drive Out Discrimination? *Paper Presented at the Annual Meeting of the American Sociological Association*.

Flage, A. (2018). Ethnic and Gender Discrimination in the Rental Housing Market: Evidence from a Meta-Analysis of Correspondence Tests, 2006-2017. *Journal of Housing Economics, 41*, 251-273.

Galster, G. C. (1992). Research on Discrimination in Housing and Mortgage Markets: Assessment and Future Directions. *Housing Policy Debate, 3*, 637-683.

Galster, G. C. (1996). Future Directions in Mortage Discrimination Research and Enforcement. In J. M. Goering & R. Wienk (Eds.), *Mortgage Lending, Racial Discrimination, and Federal Policy* (pp. 697-716). Washington, DC: The Urban Institute Press.

German Statistical Office. (2017). *Nationalites in Germany (Table 12411-0009)*.

Greene, W. H. (2012). *Econometric Analysis* (7th ed.). Boston: Prentice Hall.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and Using the Implicit Association Test: I. an Improved Scoring Algorithm. *Journal of Personality and Social Psychology, 85*, 197-216.

Hanson, A., Hawley, Z., & Taylor, A. (2011). Subtle Discrimination in the Rental Housing Market: Evidence from E-Mail Correspondence with Landlords. *Journal of Housing Economics, 20*, 276-284.

Hanson, A., & Santas, M. (2014). Field Experiment Tests for Discrimination against Hispanics in the US Rental Housing Market. *Southern Economic Journal, 81*, 135-167.

Harrison, M., Law, I., & Phillips, D. (2005). Migrants, Minorities and Housing: Exclusion, Discrimination and Anti-Discrimination in 15 Member States of the European Union. In: European Monitoring Centre on Racism and Xenophobia.

Heckman, J. J. (1998). Detecting Discrimination. *Journal of Economic Perspectives, 12*, 101-116.

Hellerstein, J., Neumark, D., & Troske, K. (1998). Market Forces and Sex Discrimination. *The Journal of Human Resources, 37*, 353-380.

Heylen, K., & Van den Broeck, K. (2016). Discrimination and Selection in the Belgian Private Rental Market. *Housing Studies, 31*, 223-236.

Hogan, B., & Berry, B. (2011). Racial and Ethnic Biases in Rental Housing: An Audit Study of Online Apartment Listings. *City and Community, 10*, 351-372.

Jann, B. (2014). Plotting Regression Coefficients and Other Estimates. *Stata Journal, 14*, 708-737.

Krysan, M., & Crowder, K. (2017). *The Cycle of Segregation: Social Processes and Residential Stratification*. New York: Russell Sage Foundation.

Lavallée, P., & Beaumont, J.-F. (2015). Why We Should Put Some Weight on Weights. *Survey Methods: Insights from the Field*.

Mazziotta, A., Zerr, M., & Rohmann, A. (2015). The Effects of Multiple Stigmas on Discrimination in the German Housing Market. *Social Psychology, 46*, 325-334.

Metcalf, G. (2018). Sand Castles Before the Tide? Affordable Housing in Expensive Cities. *Journal of Economic Perspectives, 32*, 59-80.
Murchie, J., & Pang, J. D. (2018). Rental Housing Discrimination Across Protected Classes: Evidence from a Randomized Experiment. *Regional Science and Urban Economics, 73*, 170-179.
Neumark, D. (2012). Detecting Discrimination in Audit and Correspondence Studies. *Journal of Human Resources, 47*, 1128-1157.
Norton, E. C., Wang, H., & Ai, C. (2004). Computing Interaction Effects and Standard Errors in Logit and Probit Models. *The Stata Journal, 4*, 154-167.
Oblom, A., & Antfolk, J. (2017). Ethnic and Gender Discrimination in the Private Rental Housing Market in Finland: A Field Experiment. *PLoS One, 12*, e0183344.
Pager, D. (2008). The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets. *Annual review of sociology, 34*, 181-209.
Pager, D., & Shepherd, H. (2008). The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets. *Annual review of sociology, 34*, 181.
Phelps, E. S. (1972). The Statistical Theory of Racism and Sexism. *American Economic Review, 62*, 659-661.
Quillian, L. (2006). New Approaches to Understanding Racial Prejudice and Discrimination. *Annual review of sociology, 32*, 299-328.
Riach, P. A., & Rich, J. (2004). Deceptive Field Experiments of Discrimination: Are they Ethical? *Kyklos, 57*, 457-470.
Ross, S., & Turner, M. A. (2005). Housing Discrimination in Metropolitan America: Explaining Changes between 1989 and 2000. *Social Problems, 52*, 152-180.
Salganik, M. (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA, US: Houghton, Mifflin and Company.
StataCorp. (2015). *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LP.
Turner, M. A., Ross, S., Galster, G. C., & Yinger, J. (2002). *Discrimination in Metropolitan Housing Markets: National Results from Phase 1 of the Housing Discrimination Study (HDS)*. Washington, DC: Urban Institute.
van Es, B., Klaassen, C. A. J., & Oudshoorn, K. (2000). Survival Analysis Under Cross-Sectional Sampling: Length Bias and Multiplicative Censoring. *Journal of Statistical Planning and Inference, 91*, 295-312.
Vuolo, M., Uggen, C., & Lageson, S. (2016). Statistical Power in Experimental Audit Studies: Cautions and Calculations for Matched Tests With Nominal Outcomes. *Sociological Methods & Research, 45*, 260-303.
Wienk, R. E., Reid, C. E., Simonson, J. C., & Eggers, F. J. (1979a). *Measuring Racial Discrimination in American Housing Markets - The Housing Markets Practices Survey*. Washington, DC: U.S. Department of Housing and Urban Development.
Wienk, R. E., Reid, C. E., Simonson, J. C., & Eggers, F. J. (1979b). *Measuring Racial Discrimination in American Housing Markets: The Housing Markets Practices Survey*. Washington, DC: U.S. Department of Housing and Urban Development.
Yinger, J. (1986). Measuring Racial Discrimination with with Fair Housing Audits: Caught in the Act. *American Economic Review, 76*, 881-893.

Appendix to:

Different Samples, Different Results?

How Sampling Techniques Affect the Results of Field Experiments on Ethnic Discrimination

Part A: Monte Carlo Simulations

Table A1. Monte Carlo Simulations for Different Sampling Strategies

Part B: Regression Results

Table A2. Multinomial Logistic Regression Models of Outcomes 'Discrimination Against the Turkish Applicant' and 'Discrimination Against the German Applicant'; Logit Coefficients and *p*-Values (in Parentheses)

Table A3. Multinomial Logistic Regressions, AMEs of Predictors and *p*-Values (in Parentheses) for the Outcomes 'Discrimination Against the Turkish Applicant' and 'Discrimination Against the German Applicant' (Reference: 'Equal Treatment') as Reported in Figure 5 (in the Main Text)

Table A4. Multinomial Logistic Regressions, AMEs for the Outcome 'Discrimination Against the Turkish Applicant' with *p*-Values (in Parentheses) for Split Samples by Supplier Size and Advertisement Time as Reported in Figure 6 (in the Main text)

Figure A1. Multinomial Logistic Regressions, AMEs for the Outcome 'Discrimination Against the German Applicant' with 95% Confidence Intervals for Split Samples by Supplier Size and Advertisement Time

Table A5. Multinomial Logistic Regressions, AMEs for the Outcome 'Discrimination Against the German Applicant' with *p*-Values (in Parentheses) for Split Samples by Supplier Size and Advertisement Time as Reported in Figure A1 (in the Appendix)

Part A: Monte Carlo Simulations

Table A1 shows the results of Monte Carlo simulations of different sampling strategies. For each sampling strategy, 50 random samples were drawn out of the market data that were collected during our 1 year observation period. The simulations are conducted for sampling intervals of one week (column 1), one month (column 2) and six months (column 3). For each interval, random samples were drawn once on the apartment level (with replacement of suppliers) and once on supplier level (without replacement of suppliers). The whole market serves as a reference category and is described in column 4.

Table A1. Monte Carlo Simulations for Different Sampling Strategies

	(1) One week		(2) One month		(3) Six months		(4) Market data
	Mean	SD	Mean	SD	Mean	SD	Mean
Sampling on apartment level:							
Advertisement time							
Mean[a]	87.74	1.30	88.14	1.20	82.38	1.31	22.19
Median[a]	60.95	1.57	60.50	1.37	58.06	1.25	13.00
Size of supplier							
Mean	1,243.02	49.60	1,286.79	65.31	1,324.49	57.25	1,246.13
Median	36.72	1.59	37.92	1.94	41.59	1.70	34.00
Observations	50		50		50		695,458
Sampling on supplier level:							
Advertisement time							
Mean[a]	70.18	3.11	70.90	4.65	59.85	5.16	22.19
Median[a]	43.99	3.60	44.20	4.71	42.33	3.44	13.00
Size of supplier							
Mean	21.48	10.20	27.42	22.89	35.86	70.61	1,246.13
Median	1.00	0.00	1.00	0.00	1.27	0.65	34.00
Observations	50		50		50		695,458

Notes. The sampling period for all simulations ended on 31st October 2015, and started the indicated time before: For column (1) on 25th October, (2) on 1st October, and (3) on 1st May 2015. For comparison, descriptive statistics on the whole market data are given (4). [a] Due to censored information on the advertisement time, we can calculate this statistic only based on 575,950 cases in the market data (see also Table 2 in the main text).

As can be seen, the sizes of suppliers in the drawn samples on the apartment level do not differ drastically from the sizes of suppliers we observe in the whole market. Sampling on supplier level, on the contrary, leads to a systematic over-representation of small suppliers—as one would expect.

For the advertisement time, we find the sampled apartments to be considerably longer advertised than we observe for the whole market. The longer the sampling interval, both for sampling on supplier level and sampling on apartment level, the shorter the advertisement time in the sample. Thus, with longer sampling intervals the gap to the observed advertisement times decreases, as one would expect. However, even for the sampling interval of six months, there remains a rather large difference to the observed advertisement times observed for the whole market.

To understand why this rather large gap between the sampled and observed advertisement times remains even for the sampling interval of six months the used sampling strategy has to be discussed in

more detail. To mirror sampling strategies used by real apartment seekers, we used a *prospective* sampling approach. On each day in each sampling period, an equal fraction of advertisements ($N_{\text{sample}}/N_{\text{days}}$) was sampled, until the intended sample size was reached. Due to sampling sequentially on each day in the field period (instead of drawing one sample for the whole field period), a length bias remains also with long sampling periods: Each day, units with longer advertisement times had a higher chance of being sampled than those with shorter advertisement times. Only when drawing *one* joint sample for the whole observation period, this sampling bias would disappear.

Thus, for apartments being advertised for longer times using a prospective strategy still results in a higher probability of being sampled.[19] However, only the prospective sampling likely matches the search strategies of real apartment seekers; and to our knowledge, this technique is also common in research so far. Using an alternative retrospective sample, i.e. drawing a sample of units that were advertised during the last week, month or even half year, there would be a very high risk that apartments are no longer available when the experiment takes place. For similar reasons, the market data used in our study would not provide an adequate sampling frame. For the length bias that exists between field experiments in contrast to real apartment seekers, therefore the comparison of different samples (and not the contrast to the market data) is of most interest. (The whole advertisement time found in the market data might, however, be used in future research as an indicator for the tightness of markets or pickiness of landlords.)

[19] Therefore, we also conducted a *retrospective* approach. With this approach, with increasing sampling periods the advertisement times approximate those in the market data at a much faster rate than in the prospective apporach. However, using the retrospective approach also strengthens the over-representation of small suppliers with longer sampling periods when sampling on the supplier level.

Part B: Regression Results

Table A2. Multinomial Logistic Regression Models of Outcomes 'Discrimination Against the Turkish Applicant' and 'Discrimination Against the German Applicant'; Logit Coefficients and *p*-Values (in Parentheses)

	M1 Only Controls		**M2 + Apartment Characteristics**		**M3[a] + Applicant Characteristics**		**M4a[a,b] + Interaction w/ Supplier Size**		**M4b[a,c] + Interaction w/ Advertisement Time**	
	Discr. T	Discr. G	Discr. T	Discr. G	Discr. T	Discr. G	Discr. T	Discr. G	Discr. T	Discr. G
Percentage of foreigners	-0.2893^{+}	0.1027	-0.2949^{+}	0.1186	-0.2954^{+}	0.1434	-0.2898^{+}	0.1395	-0.3036^{+}	0.1611
	(0.0951)	(0.6681)	(0.0918)	(0.6198)	(0.0867)	(0.5616)	(0.0839)	(0.5772)	(0.0846)	(0.5140)
City	0.0151	-0.2235	0.0334	-0.2183	0.0434	-0.2510	0.0417	-0.2470	0.0549	-0.2443
	(0.9225)	(0.3839)	(0.8285)	(0.3977)	(0.7771)	(0.3318)	(0.7825)	(0.3438)	(0.7220)	(0.3554)
Log size of supplier			-0.0830**	-0.1128	-0.0809**	-0.1111	-0.0121	-0.3330	-0.0824**	-0.1230^{+}
			(0.0077)	(0.1139)	(0.0095)	(0.1289)	(0.9170)	(0.1461)	(0.0092)	(0.0955)
Log advertisement time			-0.0519	0.0525	-0.0516	0.0556	-0.0500	0.0615	0.3875*	0.3663
			(0.2839)	(0.5389)	(0.2935)	(0.5257)	(0.3073)	(0.4686)	(0.0326)	(0.1597)
Applicant characteristics										
Educational status (ref.: high)										
No info					0.2264	0.2293	0.1475	-0.0548	0.3645	0.5660
					(0.1230)	(0.3085)	(0.3609)	(0.8411)	(0.3457)	(0.3479)
Low					0.2409^{+}	0.0445	0.1863	-0.0386	0.8573*	0.3227
					(0.0653)	(0.8592)	(0.2405)	(0.8986)	(0.0148)	(0.5807)
Employment status (ref.: employed)										
No info					0.1678	0.1410	0.2721	0.0508	0.5981	1.4620^{+}
					(0.2787)	(0.5663)	(0.1820)	(0.8705)	(0.1107)	(0.0697)
Self-employed					0.2335^{+}	0.0408	0.2217	-0.0805	0.1626	1.3064
					(0.0747)	(0.8692)	(0.1974)	(0.8069)	(0.7343)	(0.1484)
Public employee					-0.0022	-0.2674	-0.0868	-0.1913	0.3614	1.1642
					(0.9892)	(0.3415)	(0.6643)	(0.5834)	(0.3396)	(0.1739)
Interaction: log size of supplier X Applicant characteristics										
Educational status (ref.: high)										
No info							0.0766	0.3020*		
							(0.3414)	(0.0328)		
Low							0.0512	0.1221		

							0.0766	0.3020*		
Employment status (ref.: employed)										
No info							-0.1122	0.1030		
							(0.3618)	(0.5818)		
Self-employed							0.0038	0.1065		
							(0.9689)	(0.5739)		
Public employee							0.0688	-0.0488		
							(0.4894)	(0.8054)		
Interaction: log advertisement time X										
Applicant characteristics										
Educational status (ref.: high)										
No info									-0.0532	-0.1159
									(0.6624)	(0.5437)
Low									-0.2232^{+}	-0.0915
									(0.0559)	(0.6235)
Employment status (ref.: employed)										
No info									-0.1574	-0.4352^{+}
									(0.2184)	(0.0704)
Self-employed									0.0256	-0.4198
									(0.8736)	(0.1274)
Public employee									-0.1329	-0.4780^{+}
									(0.2842)	(0.0701)
Constant	-1.3750***	-3.1053***	-1.1255***	-3.1476***	-1.2380***	-3.3177***	-1.3427***	-3.2050***	-2.4843***	-4.3762***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0001)	(0.0000)
N	2,992		2,992		2,992		2,992		2,992	
AIC	3,399.86		3,396.48		3,416.91		3,436.39		3,435.32	
BIC	3,435.88		3,456.52		3,645.05		3,832.63		3,831.56	

Notes. *p*-values are in parentheses; $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. For outcome 'Discr. T' ('Discr. G'), applicant's characteristics are those of Turks (Germans). [a]Additionally controlled for: applicant's family status (with partner, with family, vs. single), and all characteristics of the applicant of the other ethnicity who e-mailed the same supplier. [b]Also controlled for the interaction of log size of supplier and applicant's family status as well as all characteristics of the applicant of the other ethnicity. [c]Also controlled for the interaction of log advertisement time and applicant's family status as well as all characteristics of the applicant of the other ethnicity.

Table A3. Multinomial Logistic Regressions, AMEs of Predictors and *p*-Values (in Parentheses) for the Outcomes 'Discrimination Against the Turkish Applicant' and 'Discrimination Against the German Applicant' (Reference: 'Equal Treatment') as Reported in Figure 5 (in the Main Text)

	Discr. T	Discr. G
Percentage of foreigners	-0.0368^{+}	0.0069
	(0.0825)	(0.4610)
City	0.0067	-0.0094
	(0.7194)	(0.3234)
Log size of supplier	-0.0092*	-0.0036
	(0.0157)	(0.1864)
Log advertisement time	-0.0066	0.0023
	(0.2716)	(0.4713)
Applicant characteristics		
Educational status (ref.: high)		
No info	0.0277	0.0087
	(0.1129)	(0.2929)
Low	0.0287^{+}	0.0017
	(0.0590)	(0.8425)
Employment status (ref.: employed)		
No info	0.0181	0.0059
	(0.3323)	(0.5370)
Self-employed	0.0265^{+}	0.0022
	(0.0957)	(0.8103)
Public employee	-0.0020	-0.0092
	(0.9132)	(0.3046)
N	2,992	
AIC	3,416.91	
BIC	3,645.05	

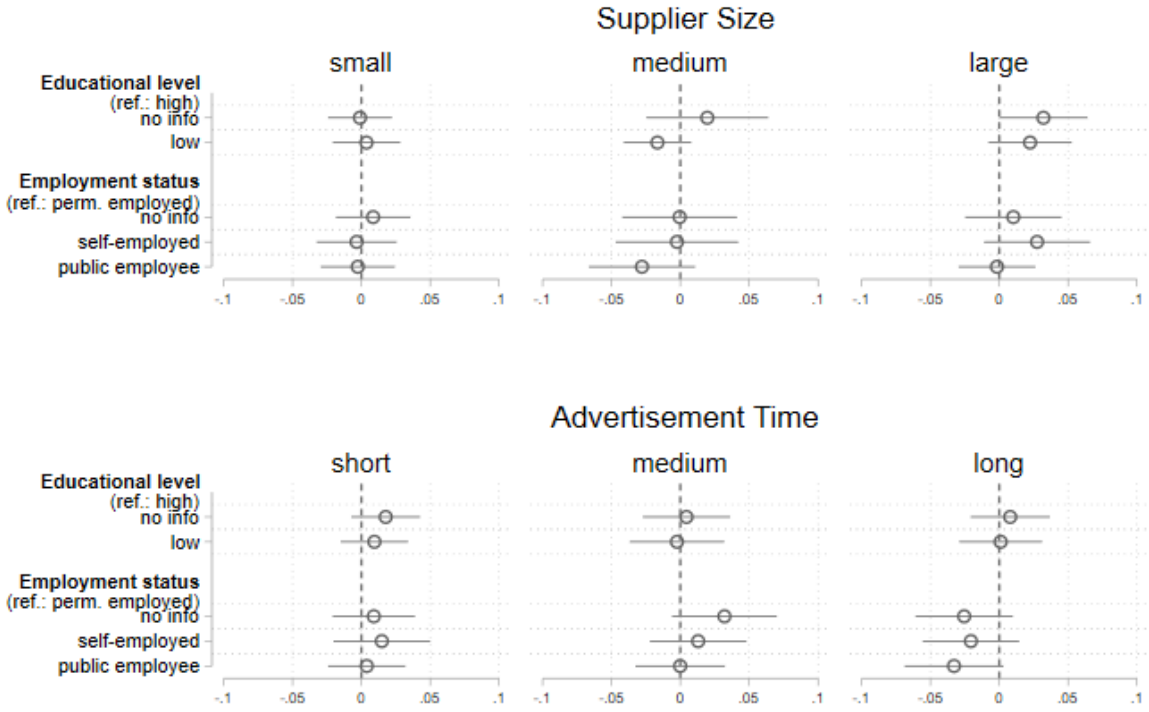
Notes. *p*-values are in parentheses; $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. For outcome 'Discr. T' ('Discr. G'), applicant's characteristics are those of Turks (Germans). Additionally controlled for: applicant's family status (with partner, with family, vs. single), and all characteristics of the applicant of the other ethnicity who e-mailed the same supplier.

Table A4. Multinomial Logistic Regressions, AMEs for the Outcome 'Discrimination Against the Turkish Applicant' with *p*-Values (in Parentheses) for Split Samples by Supplier Size and Advertisement Time as Reported in Figure 6 (in the Main text)

	Supplier Size			Advertisement Time		
	small	medium	large	short	medium	long
Percentage of foreigners	-0.0232	-0.0225	-0.0740*	-0.0317	-0.0480^{+}	-0.0224
	(0.4232)	(0.3816)	(0.0210)	(0.2585)	(0.0917)	(0.5579)
City	-0.0185	0.0118	0.0531^{+}	-0.0302	0.0375	0.0030
	(0.4809)	(0.7175)	(0.0867)	(0.3080)	(0.2148)	(0.9268)
Applicant characteristics						
Educational status (ref.: high)						
No info	0.0086	0.0574^{+}	0.0458	0.0326	0.0454^{+}	0.0025
	(0.6933)	(0.0578)	(0.1607)	(0.2099)	(0.0946)	(0.9349)
Low	0.0216	0.0286	0.0369	0.0774**	0.0252	-0.0228
	(0.3495)	(0.4173)	(0.2039)	(0.0010)	(0.3390)	(0.4348)
Employment status (ref.: employed)						
No info	0.0284	0.0235	-0.0101	0.0681*	-0.0146	-0.0027
	(0.3229)	(0.4938)	(0.7911)	(0.0115)	(0.6721)	(0.9326)
Self-employed	0.0281	0.0453	0.0137	0.0742**	-0.0118	0.0164
	(0.2488)	(0.1866)	(0.7034)	(0.0099)	(0.7199)	(0.5855)
Public employee	-0.0232	0.0414	0.0215	0.0349	-0.0281	-0.0248
	(0.3350)	(0.2504)	(0.5362)	(0.1978)	(0.3985)	(0.4066)
N	1,661	658	673	1,044	980	968
AIC	2,057.62	718.66	708.55	1,210.51	1,162.51	1,105.55
BIC	2,241.73	871.29	861.95	1,378.84	1,328.68	1,271.30

Notes. *p*-values are in parentheses; $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Displayed are AMEs estimated by separate regressions for small (only one offer), medium (2–17 offers) and large suppliers (at least 18 offers); and short (1–12 days), medium (13–29 days) and long advertisement times (at least 30 days) until treatment. Additionally controlled for: applicant's family status (with partner, with family, vs. single), and all characteristics of the applicant of the other ethnicity who e-mailed the same supplier.

Figure A1. Multinomial Logistic Regressions, AMEs for the Outcome 'Discrimination Against the German Applicant' with 95% Confidence Intervals for Split Samples by Supplier Size and Advertisement Time



Notes. Displayed are AMEs estimated by separate regressions for small (only one offer), medium (2–17 offers) and large suppliers (at least 18 offers); (top panel); and short (1–12 days), medium (13–29 days) and long advertisement times (at least 30 days) until treatment (below panel). Additionally controlled for: applicant's family status (with partner, with family, vs. single), and all characteristics of the applicant of the other ethnicity who e-mailed the same supplier, as well as the percentage of foreigners in the area and whether the apartment was located in a city yes or no. See Table A5 in the Appendix for the AMEs displayed. All estimates are based on at least N = 658 observations.

Table A5. Multinomial Logistic Regressions, AMEs for the Outcome 'Discrimination Against the German Applicant' with *p*-Values (in Parentheses) for Split Samples by Supplier Size and Advertisement Time as reported in Figure A1 (in the Appendix)

	Supplier Size			Advertisement Time		
	small	medium	large	short	medium	long
Percentage of foreigners	0.0041	0.0153	0.0086	0.0142	0.0006	0.0036
	(0.7333)	(0.2748)	(0.6287)	(0.2374)	(0.9697)	(0.8156)
City	-0.0031	-0.0276	-0.0077	-0.0087	-0.0329^{+}	0.0126
	(0.8281)	(0.1581)	(0.6946)	(0.6257)	(0.0533)	(0.4716)
Applicant characteristics						
Educational status (ref.: high)						
No info	-0.0010	0.0196	0.0321^{+}	0.0177	0.0045	0.0080
	(0.9314)	(0.3856)	(0.0503)	(0.1663)	(0.7809)	(0.5878)
Low	0.0037	-0.0166	0.0224	0.0095	-0.0024	0.0010
	(0.7671)	(0.1852)	(0.1473)	(0.4480)	(0.8904)	(0.9461)
Employment status (ref.: employed)						
No info	0.0086	-0.0005	0.0102	0.0090	0.0320^{+}	-0.0255
	(0.5354)	(0.9809)	(0.5662)	(0.5555)	(0.0994)	(0.1578)
Self-employed	-0.0034	-0.0024	0.0275	0.0149	0.0130	-0.0206
	(0.8165)	(0.9159)	(0.1621)	(0.4072)	(0.4678)	(0.2502)
Public employee	-0.0026	-0.0278	-0.0016	0.0039	-0.0001	-0.0328^{+}
	(0.8504)	(0.1582)	(0.9088)	(0.7849)	(0.9970)	(0.0738)
N	1,661	658	673	1,044	980	968
AIC	2,057.62	718.66	708.55	1,210.51	1,162.51	1,105.55
BIC	2,241.73	871.29	861.95	1,378.84	1,328.68	1,271.30

Notes. *p*-values are in parentheses; $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Displayed are AMEs estimated by separate regressions for small (only one offer), medium (2–17 offers) and large suppliers (at least 18 offers); and short (1–12 days), medium (13–29 days) and long advertisement times (at least 30 days) until treatment. Additionally controlled for: applicant's family status (with partner, with family, vs. single), and all characteristics of the applicant of the other ethnicity who e-mailed the same supplier.