From Lost Letters to Conditional E-Mail Responses: A Method to Assess Biased Behavior Online

Schott, Malte; Bluemke, Matthias

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

International Journal of Internet Science
2015, 10 (1), 49–69 ISSN 1662-5544

International Journal of Internet Science

IJIS.NET

From Lost Letters to Conditional E-Mail Responses: A Method to Assess Biased Behavior Online

Malte Schott[1], Matthias Bluemke[2]

[1]*University of Heidelberg, Germany;* [2]*GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany*

Abstract: This article introduces the *Conditional E-mail Response Technique* (CERT) as a systematic, hidden observation technique to measure behavioral tendencies. Although CERT derives from older techniques such as lost-letter/lost-e-mail techniques, we show how CERT is unique: each participant receives several e-mails with varying content, allowing the researcher to observe response rates and valence as a function of the manipulated content. Our study investigated discrimination against foreigners in the apartment rental market in Heidelberg (a German university city) by recording lessors' (non-) responses to 600 e-mails from fake applicants. Each owner (*N* = 120) received five applications for a one-room apartment via e-mail. Applicants' ethnic identities were communicated through their names. The results showed a remarkable bias against foreign names compared to German names. The response rates for foreign applicants were almost half that for German applicants (response rates were 78% for German names compared to 44–54% for American, Italian, Russian, and Turkish names). The relative risk of a rejecting response was up to eight times higher for e-mails appearing to come from foreigners. Applicants with foreign names were noticeably more likely to receive either no response or a negative response, that is, to have a negative outcome. There were also differences among the foreign applicant groups. We discuss the implications, ethical considerations, and advantages of CERT compared to other related techniques, as well as possible future uses.

Keywords: Discrimination, prejudice, lost letter, lost e-mail, ethnic minority

Introduction

Scientists have traced the pervasive influences of prejudice, stereotypes, and discrimination through contemporary times. Current issues of discrimination center around sexism/genderism (Tebbe, Moradie, & Ege, 2014), ageism (Malinen & Johnston, 2013), and weightism (Bento, White, & Zacur, 2012); discrimination also occurs within the healthcare system (Shavers et al., 2012) and against stigmatized patient groups (e.g., those with HIV; Nöstlinger, Castro, Platteau, Dias, & Le Gall, 2014). With this article, we introduce and apply a new research method that is applicable to a wide variety of discrimination issues.

Blatant and subtle forms of ethnic discrimination have been documented in various markets, including the job and consumer markets (Pager & Shepherd 2008). Empirical assessment of behavioral bias typically requires unobtrusive measures (Webb, Campbell, Schwartz, & Sechrest, 1966). Recently developed tools for uncovering cognitive or behavioral bias include implicit measurement procedures (e.g., Rudman & Ashmore, 2007; for a critical view see Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). The Internet offers social scientists a

Address correspondence to Malte Schott, Psychological Institute, University of Heidelberg, Hauptstraße 47–51, 69117 Heidelberg, Germany, Phone: +49 (0) 6221 547700, malte.schott@psychologie.uni-heidelberg.de

unique set of digital approaches to research. The present article introduces the *Conditional E-Mail Response Technique (CERT)*, a covert observation method in the tradition of the lost-letter technique (Merrit & Fowler, 1948; Milgram, Mann, & Harter, 1965). To illustrate its usability, we investigate bias in the apartment rental market as a proof of concept.

Discrimination and the housing market

In accordance with other authors, we define discrimination as an overt behavior that favors one group over another without good reason (cf. Brendl, Markman & Messner, 2001; Fiske, 1998; Brewer, 1994). The apartment rental market is a social domain where discrimination of minorities can be expected. Housing is often scarce. Because of an absence of transparency, apartment owners are under no social pressure to conform to norms of fairness or political correctness. The contract requires the lessor to interact and engage in an enduring business relationship with the lessee. Discrimination between social groups does not have to be rooted in antipathy or prejudice. "Discrimination may have causes other than prejudice" (Dovidio & Gaertner, 1986, p. 3). The possible reasons for discrimination or lessors' motivations are not the focus of this contribution. From the perspective of rental applicants, it is irrelevant *why* one does or does not achieve the desired outcome, that is, the chance to rent an apartment. Nonetheless, different ethnic groups' success in renting an apartment is a suitable measure of *perceived* behavioral bias.

Overview of hidden observation methods

Researchers have often used hidden observation to circumvent reactivity when studying behavior in socially sensitive domains (Sechrest & Below, 1983). For a variety of reasons, in an overt study, self-reported data are often unlikely to reflect biased behavior. Rather than acting as usual, individuals frequently alter their behavior to be more socially acceptable (McConahay, 1983; Pettigrew & Meertens, 1995). Hence, global and specific attitude measures may correspond imperfectly to actual behavior (Weigel & Newman, 1976). However, the problem of attitude-behavior correspondence is alleviated if researchers *covertly* observe the behavior of interest.

Before we introduce CERT as a new method for covertly observing Internet-based behavior, to properly evaluate its advantages, we briefly review previous hidden observation techniques that pertain to CERT and its application to the apartment rental market. All the presented methods have a common feature: the researcher establishes contact with a participant who is unaware of participating in a scientific investigation. Some of the information exchanged during the staged contact is misleading, deceiving the participant into responding to a supposedly genuine message that is in fact a mode of communication (or communication channel) supervised by a researcher.

Lost Letter Technique. The seminal method in this field is the well-known *lost-letter technique* (Merrit & Fowler, 1948; Milgram, Mann, & Harter, 1965). The lost-letter technique varies the recipients (e.g., organizations) of stamped but un-mailed letters, apparently lost in public by their senders. By analyzing the return rates of letters to different recipients, researchers infer the public's attitudes towards the various recipients. For instance, significant differences emerged in the rates at which letters addressed to the Friends of the Communist Party or Friends of the Nazi Party were returned (≈ 25%) compared to the rates at which letters addressed to "Medical Research Associates" or a mere private person were returned (> 70%). From this measure, researchers infer that attitudinal differences, bias, or discrimination is present.

Lost E-Mail Technique. Using time- and cost-effective electronic mail, Stern and Faber (1997) adapted the lost-letter technique for the digital age, terming their approach the *lost-e-mail technique*. In this method, the researcher sends each participant a purported e-mail *reply* that is seemingly authenticated by the inclusion of the thread of previously exchanged e-mails. The e-mail is not "lost" but rather "misdirected" (Howitt et al., 1977). Though this approach is similar to the lost-letter technique, the plausibility of receiving a misdirected e-mail may be somewhat lower from the recipient's perspective than receiving misdirected or lost paper mail. The plausibility problem can be reduced by including a unique and important message to the purported recipient (e.g., Bushman & Bonacci, 2004; Franzen & Pointner, 2013; Penney & Lawsin, 2013; Tykocinski & Bareket-Bojmel, 2009; Vaes, Paladino, & Leyens, 2002; Vaes, Castelli, Paladino, Leyens, & Giovannazi, 2003; Webb, 2011). Still, the plausibility issue limits the use of the lost-e-mail technique to between-subject designs.

Fictitious Applicant Identity Technique. Another type of covert observation is the *fictitious applicant technique*, which has been used to assess labor market discrimination. Researchers manipulate the ethnic background of purported job searchers while holding their skill levels constant, for instance by randomly assigning either an African-American-sounding name or a European-American-sounding name (Bertrand & Mullainathan, 2003) or by attaching to a resume a photograph of either a white or a black person (McConahay, 1983). Bertrand and Mullainathan quantified the response rates of 1,300 employers by counting the interview callbacks their fictitious applicants received on answering-machines. Discrimination against minority group members was reflected in

minority applicants receiving fewer callbacks irrespective of their gender, resume quality, occupation, and geographic location. This method allows a within-participant design, which strengthens causal inferences, if only at the level of the experimental unit (employers).

(Online) Mystery Research. A closely related approach is (online) mystery research, such as mystery shopping (Morrison, Colman, & Preston, 1997). The idea is to test service quality from the perspective of a business costumer, though psychiatric and other services have also been targets (e.g., walk-in pseudo-patients; Rosenhan, 1973; Lazarus, 2009). Researchers have targeted characteristics of services providers including style/formality, amount and quality of service, and accuracy and speed of reply. Unlike face-to-face interaction with shop staff, *online* mystery research methods—and mystery *e-mails* in particular—lend themselves to a higher degree of standardization (Morrison et al., 1997). Still, even well-trained mystery shoppers may not be hypothesis-blind during social interaction and are likely to react to the particulars of the situation. Indeed, empirical research suggests occasion specificity may be a problem with mystery research (Finn, 2007): repeated contact may involve different staff members at different times, therefore allowing inferences only at the level of system operations rather than at the level of the individual, which can weaken causal hypothesis tests.

Online Auctions. Shohat and Musch (2003) covertly observed online auction services (e.g., eBay). They created two different user accounts, one with a German-sounding name, and the other with a Turkish-sounding name. The idea was to analyze whether German or Turkish DVD sellers would receive better bids. The seller accounts and offers were identical except for the different names, which implied different ethnic backgrounds. For this business relationship, a rather short, one-time encounter, there was little evidence of ethnic discrimination. Note that this design could not control which participants actually saw the offers and whether both offers were seen by the same bidders. When some known confounds were statistically controlled for, evidence for ethnic discrimination was stronger, at least in the sense that "German" sellers received higher selling prices than "Turkish" sellers (Przepiorka, 2011).

Conditional e-mail response technique (CERT): Application to apartment rentals

Carpursor and Loges (2006) observed rental discrimination in Los Angeles as a function of ethnicity, which was conveyed by applicant names. Sending e-mail applications to lessors who advertised rental property, they varied *between-subjects* the name (and ethnic background) of purported tenants. African-American names and Arab names elicited significantly fewer positive responses than white (European) names.

Note that the literature review above suggests that by manipulating e-mail content *within-subjects*, the explanatory—and the statistical—power of a study can be enhanced drastically because there is stronger causal inference at the level of individual respondents. Assuming there are no preexisting differences between the experimental conditions, each participant's response pattern can be attributed to biased responding of the same experimental unit. Compared to a single observation per participant, a biased pattern of responses as a function of ethnicity reduces the likelihood that e-mails were simply left unnoticed or undelivered (ethnicity-specific spam filters are unlikely, but ultimately cannot be ruled out). Furthermore, it is a well-known advantage that within-subject designs require smaller sample sizes than between-subject manipulations, which require rather large samples to cancel out random differences between conditions, particularly in field studies (Charness, Gneezy, & Kuhn, 2012). Occasion specificity when repeatedly presenting different stimulus materials can also be reduced if it is possible to homogeneously establish contact within a short time frame.

Given these advantages, the basic idea of CERT is to approach the same participants repeatedly in a consistent fashion with standardized e-mail content to elicit, observe, and analyze their responses. In the context of this study of the apartment rental market, in which applicants' ethnicity is manipulated within participants by changing the senders' names across a series of e-mails, a behavioral bias is particularly evident in *non-responses.* If the mode of interacting, the e-mail texts, and the context in which the e-mail is read (location, time, mood of recipient, apartment availability) are comparable, (non-)responses from the same recipients are contingent only upon the key manipulated features of the e-mail requests they receive. As (non-)responses at the individual level are—by definition—confounded with idiosyncratic interactions between each recipient and the stimulus material, the data have to be aggregated across individuals to analyze responses at the experimentally controlled stimulus level, where conclusions regarding the causal effect of the experimental factor are meaningful. Obtaining a response does not yet imply a positive outcome (renting success); by contrast, a non-response necessarily constitutes a negative outcome for the applicant.

Thus, our goal was to investigate whether some applicant groups were more disadvantaged than others in terms of receiving significantly fewer (or fewer favorable) responses than other ethnic groups. The rationale was to

investigate a behavioral bias that puts foreigners at a disadvantage in a domain of *non*-behavior, namely *non-responding* to fictitious foreign applicants, that is otherwise difficult to observe.

The apartment rental market in a university city like Heidelberg is a suitable domain for using CERT for several reasons: (1) Lessors commonly provide e-mail addresses to potential tenants for the initial (asynchronous) contact. The large number of applicants in urban areas and the use of Internet platforms to advertise apartment rentals make e-mail a preferred mode of communication. (2) It is common to receive a substantial number of similarly worded e-mails in a relatively short period, all of which ask to visit the advertised property. Receiving nearly identical messages that mostly vary in senders' name and address is hardly suspicious. (3) In recent years, the housing situation in German university cities has worsened due to double-sized cohorts of German high-school graduates and a continuous influx of exchange students. The latter group may be particularly vulnerable as they often lack the resources, tacit cultural knowledge, flexibility, and opportunity to find accommodations. For instance, the proportion of foreigners in our target city has nearly tripled over the past 35 years, peaking at 16.6% in the year 2006 (FGW, 2006), and it is likely to continue to rise given the recent influx of refugees in Germany. The same survey showed that residents in this area are against continued immigration. Applicants with non-German names may experience the shortage of living space even more acutely, rendering this topic a fitting one for a first proof of concept. With the aim of demonstrating CERT's feasibility, we observed the responses of Heidelberg lessors to e-mails from supposed university students with foreign or German names.

Method

Overview: Design and hypotheses

Irrespective of lessors' motivations, we compared—within lessors—responses to fictitious tenants belonging to the five largest national groups among students at the target university: Germans, Turks, Italians, Russians and Americans (Universität Heidelberg, 2011). Using diverse, inconspicuous e-mail accounts, we sent each of 120 participants a set of five similar e-mails, each of which was randomly combined with an applicant's name. These names reflected different ethnic backgrounds. The order in which the e-mail from each name was sent and the sex of the applicant were counterbalanced across participants. A student experimenter recorded the response status and the valence of responses received within one week of contact. The student experimenter was not blind to the hypotheses (see discussion below).

We had three hypotheses. First, we expected apartment owners to be more likely to reply to applicants perceived to be German than to those perceived to be foreign (cf. Carpursor & Loges, 2006). Second, when an owner replied at all, we expected e-mails to be more positive when directed toward Germans rather than foreigners, with positive e-mails indicating the possibility of or suggesting a date for a visit to the apartment. Third, among the four foreigner groups, we expected applicants with Turkish and Russian names—the two factions of immigrants in Germany often referred to as "Gastarbeiter" (Turkish foreign workers) and "Aussiedler" (repatriates from Russia)—to be even more disadvantaged in terms of overall negative outcomes in comparison to applicants from smaller immigrant groups with American and Italian names. This assumption was based on the documented stereotypical biases against the Turkish community forming the largest migrant group in Germany before the fall of the iron curtain, and the repatriates from the former Soviet Union representing the largest migrant group after the wall came down (Heubrock, Voukava, & Petermann, 2008; Statista, 2015; Titzmann, Silbereisen., Mesch, & Schmitt-Rodermund, 2011; Wagner & Machleit, 1986; Worbs, Bund, Kohls & Babka von Gostomski, 2013).

Procedure and material

Participants and Recruitment. Figure 1 provides an overview of the study design (cf. Figure 2 for the detailed decisions involved in determining the study procedure). Any apartment owner from any part of Heidelberg who advertised a one-room apartment on an online platform specializing in student flats (*wg-gesucht.de*) was considered a potential participant. Every lessor who posted a listing and provided a contact e-mail address during a four-week period (March to April 2013) was included in the subject pool. We checked daily for new listings around noon. Lessors who fit the criteria received five e-mails (subject line: randomly either "your ad" or "wg-gesucht.de") from purported applicants in response to their new ad within less than an hour, seemingly during lunchtime. After obtaining the sample size required for counterbalancing ($N = 120$, see Independent Variable), the participant pool comprised 58 male and 62 female lessors as judged from the first names in e-mail addresses or signatures. Apart from the simple gender count made during the application phase, no identifying information—not even gender—was stored to fully protect anonymity (see Ethical Considerations below).

Stimulus Material: E-Mails. Each lessor received five different e-mails based on five templates that were randomly paired with the ethnically marked names for each participant (Appendix A: Table A1). Although the

shared origin of the e-mails was not obvious, technically experienced apartment owners might have recognized they came from the same IP address if they looked that information up in the e-mail header. Minor text variations (e.g., age) obscured the resemblance of otherwise similar e-mails (Appendix B; see Appendix C for German originals).

All the fictitious tenants identified as first-year university students between 18 and 20 years (the age of majority in Germany is 18). Local apartment hunters are frequently first-year students, especially shortly before a semester starts, as was the case during the data collection phase. The texts alleviated potential concerns that owners might have about cultural differences and communication problems or about financial resources among exchange students in that every student used perfect German style and offered a guaranteed payment, or financial bond, from their parents. All the (fabricated) applicants were as eloquent and financially stable as the typical German student. That grammatically correct e-mails from native and non-native individuals might be *perceived* differently cannot be ruled out. Recent trends support the feasibility of CERT regarding the future use of *identical* e-mails. Their credibility is greatly enhanced as more and more ready-made communication tools become available via the Internet. With click-and-send e-mail drafts, not only is it convenient for tenants to initiate first contact but it is also unremarkable for lessors to receive identically worded e-mails varying only in the sender's name and e-mail address.

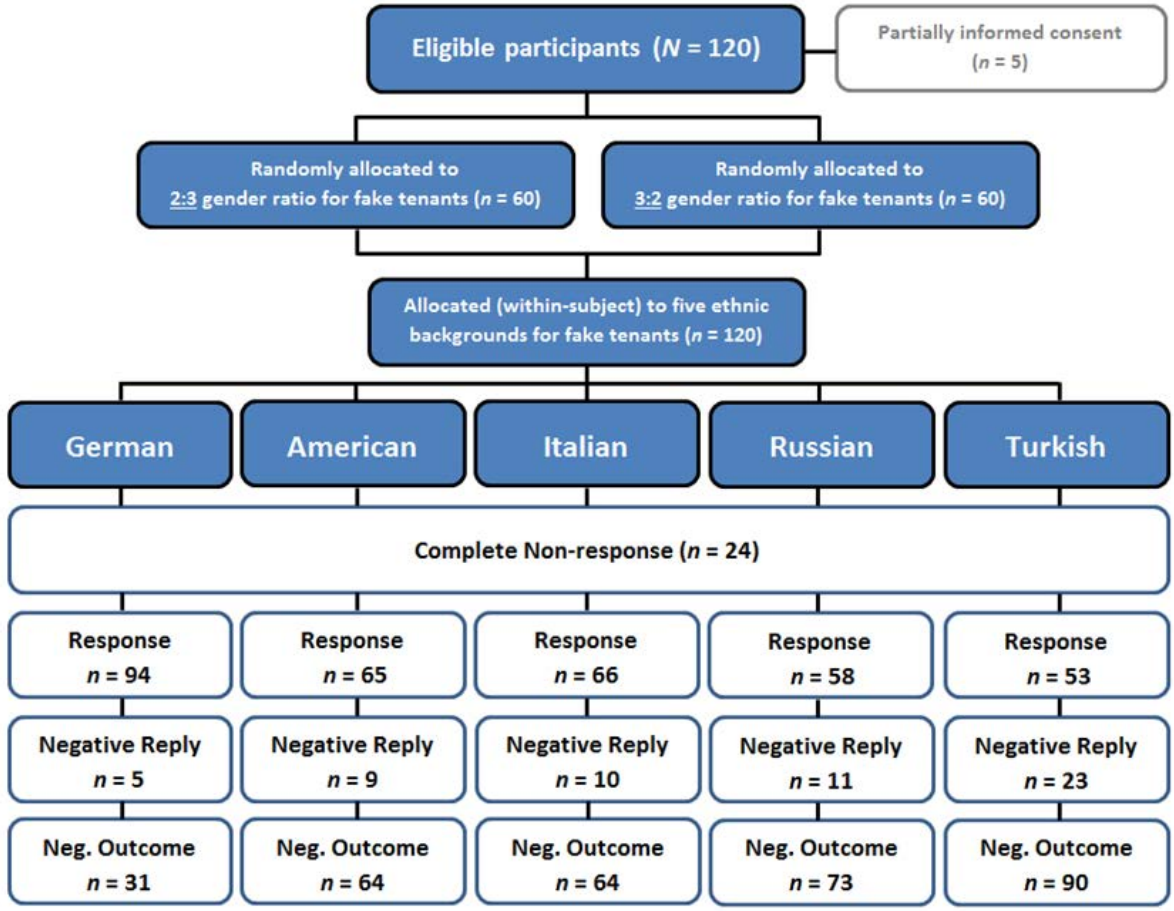

Figure 1. Flowchart: Study procedure and determination of outcome variables.

*Independent Variable: Ethnicity (Applicant Name).*The critical variation in e-mail content was the name of the sender as indicated in the signature and e-mail address (e.g., lisa.schmidt2@freenet.de, ayse_yilmaz1@gmx.de; cf. Appendix A). Different names indicated tenants' supposedly American, German, Italian, Russian, or Turkish descent. Each of the five ethnicities was represented by a first name common in the cohorts born between 1992 and 1995. To each name, we randomly assigned an e-mail provider. The set of name-provider combinations was fixed throughout the study. To verify that the intended ethnicities would be inferred correctly and that the e-mail addresses would not raise suspicion, we ran a manipulation check on the material to be used as the independent variable. The ethnic background of the name/e-mail-combinations could be identified with above 80% accuracy in a separate sample (N = 28 students who were blind to the hypotheses; Figure A1). Senders' trustworthiness was rated to be roughly comparable (Figure A2). Despite some variation in the trustworthiness scores of foreign names,

they did not differ significantly from the scores of German names. Search engine count estimates (SECEs, see Janetzko, 2008) in Google showed that the first and last names used here appear frequently on the web, with the decadic logarithm (Log10) of the page counts ranging between 7.67–9.16 for first names and 7.01–9.16 for last names.

In repeated-measurement designs, the order in which the factor levels are encountered may affect responses. We controlled for order effects by counterbalancing all possible permutations of the sender order across the included lessors. Using 5! = 120 different orders, we eliminated systematic position effects and minimized interactions with test occasions (order effects). At the same time, applicant gender was counterbalanced such that within each ethnic group, half of the e-mail senders were female. Each participant received a nearly equal 2:3 ratio of e-mails from male and female applicants (or vice versa), and the gender ratio was balanced across the full set of 600 e-mails.

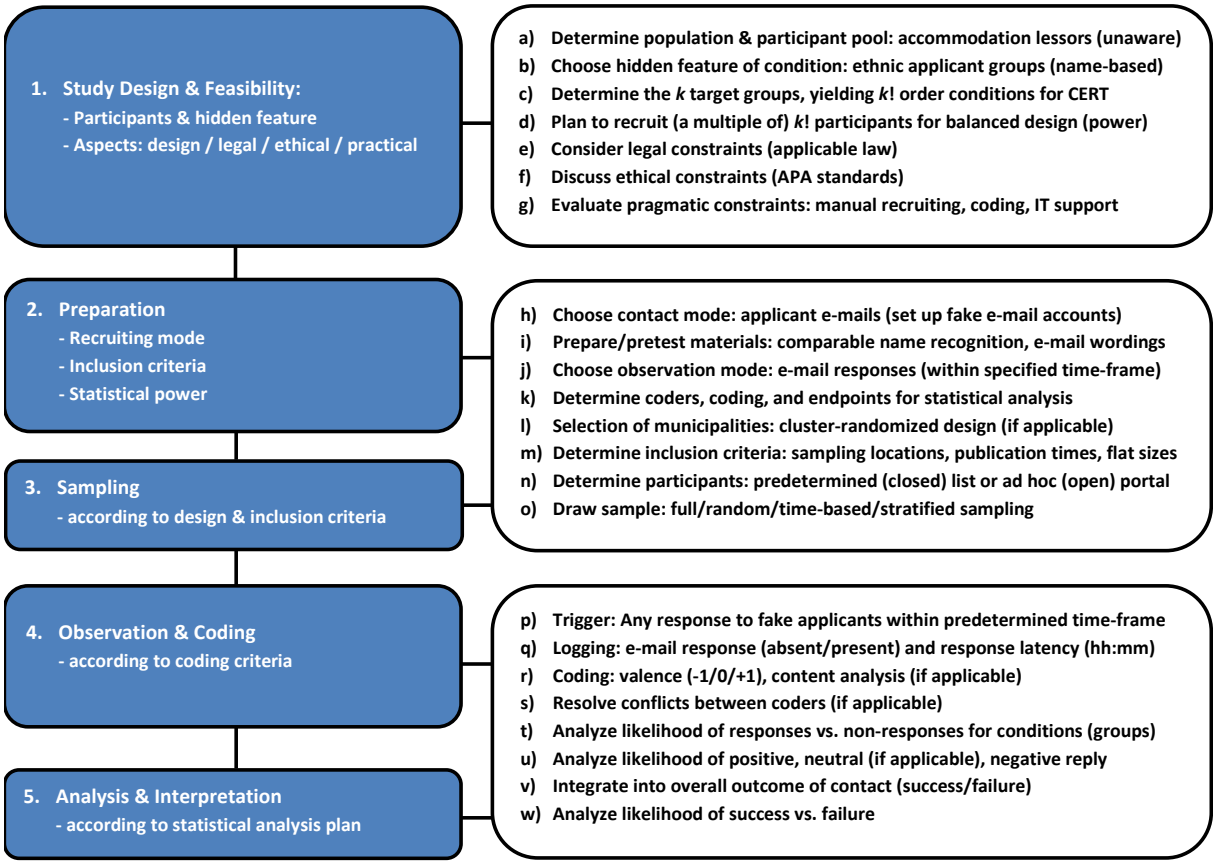

Figure 2. Flowchart: Major and minor steps and decisions of the study procedure.

*Dependent Variables: Coding of Responses for Response Rate, Valence of Reply, and Valence of Outcome.*The primary study measure (cf. Figure 1), that is, the response rate of each ethnic group, was the sum of replies to each fake applicant (1 = response, 0 = non-response). An e-mail reply was coded as present if a response was received in the relevant mailbox. All replies were received within the intended response window of seven days.

As a second outcome of interest, the student experimenter coded the valence of any response as positive or negative. If a response occurred, replies were coded as "positive" if the possibility of an appointment or a specific date to see the apartment was mentioned; it was coded as "negative" if the reply explicitly indicated that the apartment was already taken or unavailable. "Neutral" replies (e.g., "Please provide more details on your financial situation") did not exist because the request for tour was always answered either in an affirmative or negative manner. There were no ambiguous cases that would have required resolution with an independent coder.

The final criterion was the overall valence of the outcome from the applicant's perspective, that is, a positive vs. negative result of the application, the latter of which combined negative replies and non-responses. Non-responses are a negative outcome tantamount to a response indicating the flat is unavailable. In sum, the major dependent variables of the study were the overall response rate, the positive vs. negative valence of the reply, and the overall outcome from an applicant's perspective (cf. Table 1). As an additional check, we also recorded response latencies until a reply was received.

Ethical considerations

CERT requires that participants are unaware of their participation. Only with uninformed participants can one reasonably expect that participants do not alter their behavior towards applicants as a result of being observed. Therefore, informed consent, one of the ethical gold standards in psychological research (Standard 8.02; APA, 2002), cannot easily be obtained.

We tried to obtain at least partially informed consent. Independent of rental application e-mails, we invited participants with an official university letter (sent via e-mail) to participate in a study on their perceptions of the residential market and strategies for advertising. The letter contained a link to a webpage with background information on the study purpose. It was accompanied by the information that other parts of the same project were run simultaneously, implying that there might be future forms of contact. With a response rate to the online-questionnaire lower than 5% ($n = 5$), we did not analyze the data, nor do we report any questionnaire details here. As participants were reluctant to support even an inconspicuous research program, we refrained from a formal debriefing.

In any research involving deception, the research is obligated to observe the principle of proportionality to protect the reputation of participants and the confidentiality of their data (Standard 8.05; APA, 2002). For instance, minor scientific insights do not justify confronting clerks with a massive workload in addition to their regular duties. However, in the present case, the scientific insights gained outweigh the burden of five additional e-mails among the lessors' anticipated workload when dealing with tenants. Under these circumstances, researchers' responsibility to restore participants' anonymity post-hoc is pertinent. Never may any individualizing information be revealed; the anonymity of participants must never be compromised.

Finally, any relevant laws and suggestions by ethics review boards are binding. For instance, in Germany one must avoid legally binding statements in e-mails, and it is illegal to disguise one's identity from the authorities. Figure 2 presents a detailed summary of the steps involved in designing a CERT procedure.

Results

Overall response rate

Of 120 participants, 96 replied to at least one purported tenant (response rate = 80%), which demonstrates that these apartment owners were able to receive e-mails and reply. However, the frequencies of replies differed between foreign and German applicant names, warranting further analysis (Table 1).

Bias against foreign groups in (non-)responses

The first hypothesis concerned the *overall number of replies* to applicants with German-seeming names compared to applicants with foreign-seeming names. A Cochran's Q test on the binary data across the five within-subject levels showed that names indicating different ethnicities were not equally likely to elicit a response from a lessor, $\chi^2(4, N = 120) = 65.64$, $p < .001$, with $\hat{\eta}_Q^2 = .14$ suggesting a large effect size (Cohen, 1988; Serlin, Carr, & Marascuilo, 1982).

Using German names as a reference group, four 2×2 tables for dependent groups were analyzed with McNemar tests. We evaluated whether it was specifically applicants with German names who received responses more frequently than foreign applicants. To maintain a family-wise error rate of $\alpha < .05$, we used the Bonferroni-Holm multiple-testing procedure (we ordered the comparisons according to ascending p-values and verified that each p-value was less than α divided by the number of remaining tests, $\alpha^* = \alpha/4, \alpha/3, \alpha/2, \alpha$, continuing the procedure as long as the tests were significant at α^*). German applicants were more likely than applicants from each foreign group to receive replies (all ps $< .001$). As evident from the rank order of response rates and the relative risks of not obtaining a response, purported tenants with foreign names, and particularly those with Russian or Turkish names, were more disadvantaged than applicants with German names (Table 1).

Bias against foreign groups in valence of e-mail replies

Second, we expected applicants with German names to receive a larger proportion of *affirmative* (rather than *rejecting) responses* to their request to inspect the flat compared to any of the foreign applicant groups. A Cochran's Q test indicated a large significant difference across groups in the proportion of replies that were positive, $\chi^2(4, N = 38) = 32.11$, $p < .001$, $\hat{\eta}_Q^2 = .21$. Although this test generally has excellent power, in this case, it

is constrained by the regrettably small number of cases in some cells because only lessors who replied to all five groups are included in this test, and the risk of missing responses increases with the number of groups compared. Because they compare only two groups, McNemar tests suffer less from a small number of replies than do Q tests; furthermore, they can be based on exact probabilities of binomial distributions to compensate for small cell sizes. All four comparisons showed that when a lessor replied, the risk of receiving a negative reply was greater if the applicant carried a foreign name (all *p*s < .05 and significant at adjusted α*-levels; cf. Table 1). For Turkish names, the risk of a negative response was eight times higher than that for German names.

Table 1
Comparison of Applicants with Foreign Names to Applicants with German Names (Within-Participants)

	German	American	Italian	Russian	Turkish
(1) Response					
Non-response	26 (22%)	55 (46%)	54 (45%)	62 (52%)	67 (56%)
Response	94 (78%)	65 (54%)	66 (55%)	58 (48%)	53 (44%)
χ^2	-	25.29	22.78	30.63	39.02
RR	1.00	2.12	2.08	2.38	2.58
(2) Valence					
Negative reply	5 (4%)	9 (8%)	10 (8%)	11 (9%)	23 (19%)
Positive reply	89 (74%)	56 (47%)	56 (47%)	47 (39%)	30 (25%)
Exact Probability	-	.031	.016	.016	< .001
RR	1.00	2.60	2.85	3.57	8.16
(3) Outcome					
Negative outcome	31 (26%)	64 (53%)	64 (53%)	73 (61%)	90 (75%)
Positive outcome	89 (74%)	56 (47%)	56 (47%)	47 (39%)	30 (25%)
χ^2	-	29.26	27.68	36.54	57.02
RR	1.00	2.06	2.06	2.35	2.90

Note. N = 120. McNemar-tests (χ^2, *df*=1) and relative risks (risk ratios, RR) for obtaining (1) no response (χ^2-based), (2) a negative reply (based on exact binomial distribution), and (3) a negative overall outcome (χ^2-based) as a function of ethnicity.

Bias against foreign groups in valence of overall outcome

Collapsing across both negative outcomes (no response and refusal of the request to visit), we analyzed differences in the *valence of the overall outcome*. The previously observed bias re-surfaced when comparing applicants with German names and those with foreign names, $\chi^2(4, N = 120) = 98.67$, $p < .001$, $\hat{\eta}_Q^2 = .21$. McNemar tests showed that applicants with German names had more positive outcomes (rather than rejections or no response) than did any other group, whether that group was applicants with American, Italian, Russian, or Turkish names (all *p*s < .001). Applicants with Russian or Turkish names were least likely to be invited to an apartment visit, confirming the third hypothesis.

Bias within the foreign groups

To check for significant differences exclusively among the four foreign ethnic group labels, the tests were repeated for all three measures: (1) responding (vs. non-responding), (2) valence of feedback, and (3) valence of overall outcome. The first Cochran's Q test showed that foreign applicant groups differed in their likelihood of eliciting any response, albeit on a smaller scale, $\chi^2(3, N = 120) = 8.27$, $p = .04$, $\hat{\eta}_Q^2 = .03$.

However, when we tested our second hypothesis by analyzing the valence of responses sent by participants who responded to all prospective tenants, there were large discrepancies in the number of positive responses sent to the four foreign groups, $\chi^2(3, N = 38) = 20.09$, $p < .001$, $\hat{\eta}_Q^2 = .18$. When we analyzed the third measure, with non-responses and refusals combined, a Cochran's Q test confirmed a substantial difference in the rate of positive outcomes across the groups with foreign names, $\chi^2(3, N = 120) = 27.46$, $p < .001$, $\hat{\eta}_Q^2 = .08$. McNemar tests showed that applicants with Turkish names received significantly fewer positive results than did applicants with other foreign names (all *p*s < .006). Neither with nor without adjustment for multiple testing did any other group comparison reach statistical significance. Reconfirming our third hypothesis, applicants with Turkish names, the largest group of immigrants in Germany, were the most disadvantaged group among the four foreign groups.

Order and position effects

Compared to a Latin square design, a fully counterbalanced within-subject design is advantageous in that it not only controls for position effects but also minimizes the possibility that order effects create aggregate-level differences between experimental conditions. Perfect counterbalancing distributes any order effects evenly across conditions (here, 120 unique orderings), so that the experimental effect will not be confounded with order of conditions. Of course, we cannot preclude asymmetrical influences between specific sequences of ethnic names, whether presented immediately one after the other or interleaved with one, two, or three different names. For instance, responses to Turkish names might depend on whether participants encountered a German name immediately beforehand, but the reverse may not hold. Given 120 unique orderings and the additional uniqueness of ethnic background/gender-combinations, we expect the size of systematic order effects to be negligible. Note that an accurate interpretation of order effects is also contingent on lessors always interacting with e-mails exactly in the order in which the researcher sent the e-mails. This assumption may not at all be justified. Instead, some apartment owners may check their e-mails only once a day and thus face a multiple-choice task; some may encounter higher e-mail traffic than others. In any case, the likelihood of substantial carry-over effects decreases with the number of orders implemented.

Table 2
Exploratory Comparison of Applicants with Foreign Names to Applicants with German Names (Between-Participants)

	German	American	Italian	Russian	Turkish
Position 1					
Response	1.00	1.67	1.33	2.00	2.33
Valence	1.00	2.57	1.13	1.50	5.40
Outcome	1.00	1.71	1.29	1.86	2.43
Position 2					
Response	1.00	1.43	2.14	1.86	1.57
Valence	1.00	3.60	5.40	7.50	19.3
Outcome	1.00	1.57	2.29	2.14	2.57
Position 3					
Response	1.00	12.0	13.0	14.0	14.0
Valence	1.00	9.23	18.0	10.9	24.0
Outcome	1.00	14.0	17.0	16.0	19.0
Position 4					
Response	1.00	3.25	2.75	3.25	3.59
Valence	1.00	7.27	4.62	3.64	8.00
Outcome	1.00	3.40	2.80	3.00	3.60
Position 5					
Response	1.00	1.25	0.88	1.25	1.75
Valence	1.00	0.16	0.31	1.52	2.13
Outcome	1.00	0.91	0.73	1.27	1.64

Note. Relative risks (RR) for receiving no response ("response"), a negative reply ("valence"), and a negative overall outcome ("outcome") as a function of ethnicity and e-mail position (1-5). Risks are relative to the risks of the reference group (German names; RR = 1). The disproportionately high risks for position 3 represent special cases of 2×2 matrices with one extremely low number of observations (1) in one German cell.

Unlike a typical cross-over (AB/BA) design, which allows the researcher to estimate an interaction term between the experimental factor and the order of conditions, this design, with exactly one participant seeing one of the 120 orders, has no variability within orders that can be exploited to estimate the error variance of the interaction term. Although carry-over effects are unlikely here by design, if one considered them likely, multiples of $k!$ orderings would be required to properly test interaction, and statistical power would have to be considered accordingly. A workaround might be to analyze (sets of) order conditions with specific commonalities, holding constant the position of an ethnic group, for example. In this fashion, one could analyze lessors' replies to applicants with Turkish names when that e-mail was sent after (or before) the e-mail from a German sender. However, as the number of possible tests is large and the design was not intended to permit analysis of interaction effects, the number of observed cases was too low for any meaningful comparison of relative risks as a function of order. In

other words, we neither claim nor rule out the possibility that the likelihood of, say, positive replies to Turkish tenants depends, to some degree, on the ethnicity of previously encountered applicants.
Nevertheless, we can evaluate whether the positions influenced the relative risks in a (pseudo) between-subjects comparison by estimating the relative risks position-by-position. We treated an outcome variable at each of the five positions as if it contained 24 completely independent observations for each ethnic group (which, strictly speaking, is true for only position #1, before any other applicant e-mails were received). Separately for the five positions, we again calculated the risk for each foreign applicant group relative to the risk for German applicants for each of the three major outcome variables. These risk ratios, now based on data from *different* lessors (Table 2), mirror the general pattern of the risk ratios computed *within lessors* (cf. Table 1): Apart from a few exceptions, where the least disadvantaged foreign names (American and Italian names) were associated with numerically lower risks than German names, foreign names consistently bore higher risks than German names (RR > 1). Irrespective of position, the most disadvantaged group (Turkish names) was at highest risk. Russian names had the second highest risk in more than 50% of the computed risk ratios.

Bias in response latencies

Imagine a politically correct world in which all prospective tenants receive positive feedback, say, due to the success of a regulatory body. Further information on biased behavior might be available in response latencies. In other words, even when there is irrefutable evidence of non-bias in a focal variable, each reply's delay might still indicate preferences for specific applicant groups. One can analyze whether some conditions elicit positive replies more quickly than others. This analysis must necessarily disregard the (unknown) sequence in which participants opened e-mails in their inbox (e.g., oldest first, newest first) and whether the recipient's e-mail traffic was high or low at the time of the study (e.g., whether experimental e-mails were interspersed with non-experimental e-mails).

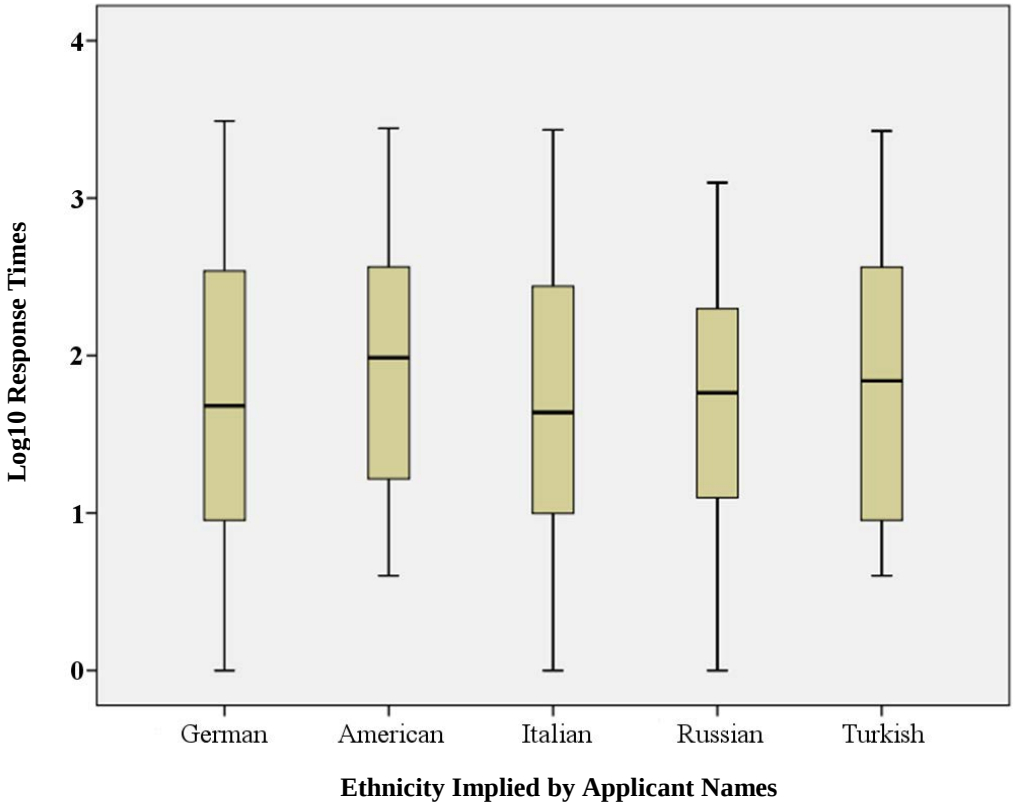

Figure 3. Box plots of Log10 response times for positive replies to the five applicant groups.

In our study, inferring preferences from any differences in latencies must be done with extreme caution because we lack the precondition of the envisioned politically correct world with a balanced design of identically valenced responses. There were only $n = 38$ complete responders, all with considerable latency variability, and the overall proportion of positive responses to purported foreign applicants differed between groups and was generally rather low (25%–47%) compared to that for German applicants (74%). Given that latencies are confounded with the decision to reply at all, we face an unbalanced design with only $n = 19$ participants responding positively to all five applicant groups. The confidence of any interpretation is undermined to the degree that the response rates are (a) low, (b) unevenly distributed across ethnicities, and (c) confounded with specific subsets of lessors who replied

differently to the fictitious applicant groups. In such a case, a paradoxical situation can emerge (Simpson, 1951): it is unpredictable whether aggregated latencies will corroborate or contradict any previously observed bias in response rates.

For exploratory purposes, we describe here the latencies for positive feedback as a function of purported ethnicity. The central tendencies (medians here, due to the skewed distributions) for positive responses towards purported German, American, Italian, Russian, and Turkish senders (*n*s = 89, 56, 56, 47, and 30) were *Md* = 48, 97, 43.5, 58, and 69.5 minutes, respectively. Disregarding the fact that the response times for each ethnic group come from different lessors and that different numbers of observations contribute to each ethnic group's data, Figure 3 displays the distributions of the Log10 response times. They show considerable overlap between all ethnic groups.

Subgroup analyses

Decision-making researchers might wonder whether lessors responded differently if they had already read all five experimental e-mails. These participants would have faced a choice situation with a full comparison of multiple (ethnic) options, and it is unclear whether our findings generalize across different decision-making situations. Note that reconstructing the decision-making process post-hoc is notoriously difficult. We do not know about relevant conditions such as (1) a lessor's overall amount of e-mail traffic, (2) the number of e-mails from real applicants (not to mention their ethnic composition), (3) the actual sequence in which lessors opened incoming e-mail, and (4) whether the order of replying mirrored the order of reading at all.

Nevertheless, we can at least determine whether it was technically possible that lessors, before they replied, had received and potentially read all experimental applications (and maybe also some from real applicants). If the minimum response time (i.e., the time of the first reply) was more than 60 minutes after the first e-mail was sent, the lessor could potentially pick from all five ethnic names those s/he liked best. Of the 96 participants who responded at least once, 41 lessors answered only *after* we had sent all five e-mails (*n* = 55 started replying *before*).

On the one hand, having seen all five e-mails might have created the impression that applications from students with foreign names are currently common and in fact outnumber German applications (in the experimental set of e-mails), thus inspiring less discrimination. On the other hand, not yet having seen all e-mails may have "accidentally" fed a bias against foreign applicants if apartment owners expected more "appealing" applications to follow. Alternatively, some might have started replying to "desirable" tenants initially but felt an affective contrast between the initial desirable and subsequent undesirable applicants.

Interestingly, those participants who had potentially seen all e-mails replied to e-mails with German names without exception. This was not the case for foreign names. Hence, we analyzed the likelihood of an unfavorable outcome for any ethnic group as a function of whether lessors had potentially encountered all e-mails. The relative risks between these subgroups did not reveal systematic differences, and the integrated overall outcome measure barely deviated from unity (cf. Table 3). On the basis of this specific comparison, we found no evidence that our core findings would not generalize across different choice situations. However, it is difficult to determine on the basis of response latencies alone how strongly the decision-making situations differed in reality.

Table 3
Descriptive Subgroup Analysis of Lessors Potentially Facing a Multiple-Choice Task

	German	American	Italian	Russian	Turkish
Response	0.27	0.64	1.53	0.98	0.72
Valence	1.94	0.88	1.09	2.31	1.35
Outcome	1.01	0.72	1.34	1.19	0.99

Note. Relative risks (RRs) for obtaining no response ("response"), a negative reply ("valence"), and a negative overall outcome ("outcome") of lessors who potentially saw all e-mails before replying (*n* = 41) in comparison to those who could not have seen all e-mails before replying (*n* = 55).

Discussion

CERT-application and findings

Overall, the application of CERT was successful and identified a bias in the apartment rental market. Our results showed substantial bias against foreign applicants as evidenced in the pattern of e-mail communication, the valence of correspondence, and the overall outcomes that applicants experienced. As expected, fake applicants

with German names received more e-mail replies than any fake applicants with foreign names. The likelihood of applicants with foreign names hearing back from a lessor depended on their specific ethnicity. If they did hear back, applicants with foreign names received affirmative feedback less frequently than applicants with German names. We obtained these outcomes despite all e-mail applications coming from broadly similar applicants with apparently identical financial security, good manners, and obviously good language skills. With type of name varying within participants and 80% of participants replying to at least one of the e-mails, technical failures can be ruled out as an alternative explanation for the obtained results.

Instead, our findings speak to the difficulties that foreigners may encounter when entering the local apartment rental market. We interpret these findings as showing that even in a social climate that deems acts of discrimination against foreigners politically incorrect, many decades after the onset of discrimination research and years after immigrants started shifting into the German education system and apartment rental market, having the "wrong" name can still put people at severe disadvantage (cf. Carpursor & Loges, 2006). This bias occurred even though we focused on accommodation in an internationally oriented, multicultural university town. Discrimination is especially likely in those areas of society that are not, or only insufficiently, transparent to the public. It is in these areas that one can discriminate against others without having to fear any negative social or legal consequences.

Although our findings support the interpretation of this bias as discrimination, the present paper is merely an initial piece of evidence pointing toward this conclusion. Whether discrimination was actually occurring cannot be established by our data as we lack further self-disclosure from the lessors regarding the reasons for their (in-)action. Therefore, each individual could have good reasons, say, based on prior experience with group members, not to consider immigrant tenants, negating an aspect of the definition of discrimination. For instance, the likelihood of having previously rented to a member of a group may be proportional to the base rate of that group. The likelihood of prior experience differs among the foreign groups, and apartment owners may simply feel more comfortable dealing with tenants whose backgrounds they are familiar with. Note that such an influence alone cannot explain why lessors would be less biased against Americans and Russians than against Turks as the latter (including their offspring) constitute the biggest faction. Apart from this, one might ask not only *if* discrimination is the case, but *who* suffers most from it, and *why*. Why is it the case that applicants with foreign names are treated as second-class tenants compared to applicants with German names? And why is there a notable discrepancy between Turks and other foreign groups? Our findings may inspire future research to answer these questions. Other domains in which biased behavior may play a similarly central role are yet to be identified.

Comparison of CERT to other hidden observation techniques

Several hidden observation techniques have been developed as technology has advanced, serving various scientific goals. CERT, with its key elements—hidden within-participant observation via the exchange of solicited e-mails—is a valuable tool for registering data on real-life behavior that may not be collected otherwise. Lessors may be hesitant to cooperate in a scientific study on discriminatory behavior, yet the behavior may still be observed covertly. The overall response rate was highly satisfactory.

Like other methods based on e-mail, CERT targets specific participants or participant groups. Although we refrained from doing so, CERT allows for tracking behavioral responses to specific individuals. The extent to which they can remain anonymous may depend on the information already available about them (e.g., from the e-mail alias). Distinct from previous approaches, the within-participant design allows testing more than two experimental conditions in an economical and inexpensive manner. The design effectively rules out that reasons other than the manipulated e-mail characteristics are responsible for any outcome differences. Other hidden observation techniques only partially lend themselves to causal inferences and typically require bigger sample sizes to do so (e.g., Carpursor and Loges, 2006).

CERT comes with the properties of a controlled experiment: (1) As soon as any fake tenant's e-mail elicits at least one response from a participant, we can rule out uncontrolled factors as an alternative explanation to non-response such as corrupted e-mail accounts, participants' absence, etc. Thus, the overall response rate gives an indication of how well the experiment was concealed. (2) Each fake tenant has *a priori* the same probability of receiving a response. The probability of being randomly drawn by a lessor is identical for the five fake applicants even if we can expect a lessor's application pool to be bigger than the five e-mails we sent. Even if 100 additional real e-mail applications were received, then the ratio for each of the five fake tenants remains identical (1:105). The probability of any fake e-mail being drawn remains equal irrespective of the number of true applicants.

The experimental manipulation—ethnicity implying names—is quite common for hidden observation in the field of discrimination research. However, CERT is suitable for a wide range of other manipulations, hence the generic name *Conditional E-Mail Response Technique*. CERT enables researchers to standardize the content of e-mails

and peripheral variations in the stimulus material to a high degree. The plausibility of receiving similar e-mails is especially high for the apartment rental market. Lessors invite e-mail contact themselves and will not become suspicious when they receive a number of similarly worded e-mails. CERT is particularly suited for domains where such circumstances are met, yet it may also be feasible to obtain valid data when contact is not explicitly encouraged by participants, as long as establishing contact via e-mail is not suspicious in itself.

Extending the work by Carpursor and Loges (2006), we used five levels for the within-subject ethnicity factor. More levels (and more factors) are possible as long as plausibility is not undermined. In comparison to other hidden observation techniques that use between-subject designs, the higher statistical power of CERT is advantageous. Obviously, this first proof of concept opens the door for a variety of other applications and as such is not limited to the domain of apartment rental markets and respective Internet platforms. The scope of CERT clearly extends beyond the domain of discrimination. Imagine a non-profit organization testing different ways of targeting e-mails to elicit donations more effectively. As another example, not only individual recipients but complete organizational systems can be examined with CERT, as with online customer care, where individuals cannot be tracked and e-mails are handled by a call center or multiple back-office assistants.

Limitations

Despite CERT being a cost-effective way to covertly observe real-life interactions, a within-participant design has specific limitations. Though we ran a fully permutated (order × target group) design, order effects can nonetheless inflate overall variance and reduce statistical power. Even worse, without a sufficient sample size, systematic interactions may go unnoticed and spuriously exacerbate or attenuate the differences between conditions. One solution is to increase the number of factor levels (names) to reduce the relative impact of specific sequences (still assuming a fully counterbalanced design), yet this approach may conflict with the feasibility of the design or with plausibility on the recipients' side. Another solution is to use an unbalanced design with sequences that are known, or at least likely, to be unproblematic.

From an experimenter's point of view, all these considerations are important but somewhat beside the point. As a method for field experiments, CERT lacks the degree of control that laboratory experiments offer. As long as one does not have full control over where, when, how, and in which order participants eventually receive and interact with e-mails (including those sent by independent true applicants), the order factor on the researcher's side will merely approximate the order of conditions on the participant's side. We could not control how many true applications an apartment owner received, what type of and how much information other e-mails may have provided, or whether those applicants revealed their ethnic background or specifically concealed it with innocuous e-mail aliases. Another limitation in our case, then, is that neither the full absence of responses nor severe delays or timely responses nor the relative timing of positive and negative responses can be causally linked to cognitive factors underlying the individual decision-making processes. Inferences should be drawn from the unsupervised parts of the communication process only with great caution. This is why we stress that the strength of CERT lies at the aggregate level of stimulus analyses, that is, the manipulated factor levels.

Several interesting topics cannot be addressed with data generated by CERT. They mostly concern individual participants' behavior and underlying motivations. From CERT, we cannot learn the reasons why lessors are less likely to rent out their apartments to Turkish applicants. We cannot even conclude whether lessors are *aware* that their own behavior is discriminatory or whether they are aware that a bias exists across the decisions of many lessors. If one e-mail was answered by a respondent, we may reasonably assume that the e-mail account was working properly, yet what do we conclude if a participant did not respond at all? We might suspect there was a typo leading to "misdirected" e-mails, but we could just as easily infer that this participant was unwilling to rent to all the tenant groups we implemented. Imagine an apartment owner who has already rented out some flats to foreign students and may from now on prefer German applicants or exchange students from a group not included in our design (say, Spanish). This example shows that it would be unjustified to deduce prejudiced behavior at the individual level as the apartment owner in fact has acted in the least prejudiced manner previously. In the absence of background information on the participants themselves, CERT is not biased by individual motivations, but it can also never reveal the individual motivations to us. No method can supply all the answers; other methods must be considered, too, when approaching these questions and testing assumptions about the causal processes underlying the phenomena observed by CERT.

A final limitation of our study is that coding was carried out by a single coder who was not blind to the hypotheses, so intercoder-reliability could not be determined. This appeared permissible after initial checks and random checks convinced us that the coding scheme was easy and data entry performed correctly. However, other coding schemes may be more complex than merely recording response vs. non-response and affirmative vs. declining reply. For instance, participants' e-mail replies may contain enough content to be coded according to linguistic categories;

alternatively, statements may need to be scored (rated) subjectively by coders for other criteria (Fiedler, Bluemke, Friese, & Hofmann, 2003). In such situations, two or more coders, who should ideally be blind to hypotheses, would be advisable. Freelon (2013) developed a helpful tool for computing the intercoder reliability online (the tool can be accessed at http://dfreelon.org/utils/recalfront/recal-oir).

Conclusion

We demonstrated that CERT is a useful approach to assess behavioral bias, especially when socially disapproved behavior is involved and becomes evident mostly by non-responses. We illustrated its utility by documenting a bias against foreigners who look for flats on the apartment rental market, a bias that may account for perceived discrimination. We did not discuss all features that researchers may want to consider before ultimately settling on their design (e.g., overall response rates, availability of technology; see Shih & Fan, 2007). Although not unaware of its limitations, we are confident that by pointing out advantages of a controlled within-participant e-mail design with relatively strong causal inferences regarding the crucial independent variable, we can inspire other researchers to apply this efficient procedure, particularly when data on behavioral bias cannot be gathered otherwise.

Acknowledgements

Rebecca Ajnwojner's assistance with the development of materials and with data collection is gratefully acknowledged, as is the editorial guidance by Ulf-Dietrich Reips and the feedback provided by the reviewers of previous versions of this manuscript, and Anastasia Penner's help with preparing the figures.

References

APA – American Psychological Association (2010). *Ethical principles of psychologists and code of conduct (with the 2010 amendments).* Retrieved from http://www.apa.org/ethics/code/principles.pdf

Bento, R. F., White, L. F., & Zacur, S. R. (2012). The stigma of obesity and discrimination in performance appraisal: A theoretical model. *International Journal of Human Resource Management, 23*, 3196–3224. doi:10.1080/09585192.2011.637073

Bertrand, M., & Mullainathan, S. (2003). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor. *NBER Working Paper* No. 9873.

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology, 81,* 760–773. doi:10.1037/0022-3514.81.5.760

Brewer, M. B. (1994). The social psychology of prejudice: Getting it all together. In M. P. Zanna & J. M. Olsen (Eds.), *The Psychology of Prejudice: The Ontario Symposium* (Vol. 7, pp. 315–329). Hillside, NJ: Erlbaum.

Bushman, B. J., & Bonacci, A. M. (2004). You've got mail: Using e-mail to examine the effect of prejudiced attitudes on discrimination against Arabs. *Journal of Experimental Social Psychology, 40*, 753–759. doi:10.1016/j.jesp.2004.02.001

Carpursor, A. G., & Loges, W. E. (2006). Rental discrimination and ethnicity in names. *Journal of Applied Social Psychology, 36*, 934–952. doi:10.1111/j.0021-9029.2006.00050.x

Charness, G., Gneezy, U., & Kuhn, M. A. (2012): Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior and Organization, 81,* 1–8. doi:10.1016/j.jebo.2011.08.009

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dovidio, J. F., & Gaertner, S. L. (1986). Prejudice, discrimination, and racism: Historical trends and contemporary approaches. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 1–34). San Diego, CA, US: Academic Press.

Fiedler, K., Bluemke, M., Friese, M., & Hofmann, W. (2003). On the different uses of linguistic abstractness. From LIB to LEB and beyond. *European Journal of Social Psychology, 33*, 441–453. doi:10.1002/ejsp.158

Finn, A. (2007). Doing a double take: Accounting for occasions in service performance assessment. *Journal of Service Research, 9*, 372–387. doi:10.1177/1094670507301065

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, and G. Lindzey (Eds.), *Handbook of Social Psychology* (4th ed., Vol. 2, pp. 357–411). New York, NY: McGraw-Hill.

FGW – Forschungsgruppe Wahlen Mannheim (2006). Demografischer Wandel in Heidelberg: Einstellungen und Meinungen zur Integration und zur weiteren Zuwanderung von Ausländern. [Demographic change at Heidelberg: Attitudes and opinions about the integration of immigrants and further in-migration of foreigners.] *Statistisches Monatsheft Baden-Württemberg 9/2006.* Retrieved from http://www.statistik.baden-wuerttemberg.de/veroeffentl/Monatshefte/PDF/Beitrag06_09_03.pdf

Franzen, A., & Pointner, S. (2013). The external validity of giving in the dictator game. A field experiment using the misdirected letter technique. *Experimental Economics, 16,* 155–169. doi:10.1007/s10683-012-9337-5

Freelon, D. (2013). ReCal OIR: Ordinal, interval, and ratio intercoder reliability as a web service. *International Journal of Internet Science, 8*, 10–16.

Heubrock, D., Voukava, L., & Petermann, F. (2008). Sind Aussiedler aggressiver? Ein empirischer Vergleich zwischen delinquenten und nichtdelinquenten Deutschen und Russlanddeutschen. [Are repatriates more aggressive? An empirical analysis of delinquent and nondelinquent Germans and ethnic German immigrants]. *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie, 56*, 293–299. doi: 10.1024/1661-4747.56.4.293

Howitt, D., Craven, G., Iveson, C., Kremer, J., McCabe, J., & Rolph, T. (1977). The misdirected letter. *British Journal of Social and Clinical Psychology, 16,* 285–286.

Janetzko, D. (2008). Objectivity, reliability, and validity of search engine count estimates. *International Journal of Internet Science, 3*, 7–33.

Lazarus, A. (2009). Improving psychiatric services through mystery shopping. *Psychiatric Services*, *60*, 972–973. doi:10.1176/appi.ps.60.7.972

McConahay, J. B. (1983). Modern racism and modern discrimination: The effects of race, racial attitudes, and context on simulated hiring decisions. *Personality and Social Psychology Bulletin, 9*, 551–558. doi:10.1177/0146167283094004

Malinen, S., & Johnston L. (2013). Workplace ageism: Discovering hidden bias. *Experimental Aging Research, 39*, 445–465. doi:10.1080/0361073X.2013.808111

Merritt, C. B., & Fowler, R. G. (1948). The pecuniary honesty of the public at large. *Journal of Abnormal and Social Psychology, 43*, 90–93. doi:10.1037/h0061846

Milgram, S., Mann, L., & Harter, S. (1965). The lost-letter technique: A tool of social research. *Public Opinion Quarterly, 29*, 437–438.

Morrison, L. J., Colman, A. M., & Preston, C. C. (1997). Mystery customer research: Cognitive processes affecting accuracy. *Journal of the Market Research Society, 39*, 349–361.

Nöstlinger, C., Castro, D. R., Platteau, T., Dias, S., & Le Gall, J. (2014). HIV-related discrimination in European health care settings. *AIDS Patient Care and STDs, 28*, 155-161. doi:10.1089/apc.2013.0247

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology, 105*, 171–192. doi:10.1037/a0032734

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology*, *34*, 181–209. doi:10.1146/annurev.soc.33.040406.131740

Penney, E., & Lawsin, C. (2013). Application of the MODE model to implicit weight prejudice and its influence on expressed and actual behavior among college students. *Journal of Applied Social Psychology, 43*(Suppl 2), 229–236. doi:10.1111/jasp.12031

Pettigrew, T. F., & Meertens, R. W. (1995). Subtle and blatant prejudice in Western Europe. *European Journal of Social Psychology 25*, 57–75. doi:10.1002/ejsp.2420250106

Przepiorka, W. (2011). Ethnic discrimination and signals of trustworthiness in an online market: evidence from two field experiments. *Zeitschrift für Soziologie, 40,* 132–141.

Rosenhan, D. L. (1973). On being sane in insane places. *Science, 179*, 250–258. doi:10.1126/science. 179.4070.250

Rudman, L. A., & Ashmore, G. D. (2007). Discrimination and the Implicit Association Test. *Group Processes Intergroup Relations, 10*, 359–372. doi:10.1177/1368430207078696

Sechrest, L., & Below, J. (1983). Nonreactive measures of social attitudes. *Applied Social Psychology Annual, 4*, 23–64.

Serlin, R. C., Carr, J., & Marascuilo, L. A. (1982). A measure of association for selected nonparametric procedures. *Psychological Bulletin, 92*, 786–790. doi:10.1037/0033-2909.92.3.786

Shavers, V. L., Fagan, P., Jones, D., Klein, W. M. P., Boyington, J., Moten, C., & Rorie, E. (2012). The state of research on racial/ethnic discrimination in the receipt of health care. *American Journal of Public Health, 102*, 953–966. doi:10.2105/AJPH.2012.300773

Shih, T.-H., & Fan, X. (2007). Response rates and mode preferences in web-mail mixed-mode surveys: A meta-analysis. *International Journal of Internet Science, 2*, 59–82.

Shohat, M., & Musch, J. (2003). Online auctions as a research tool: A field experiment on ethnic discrimination. *Swiss Journal of Psychology, 62*, 139–145. doi:10.1024//1421-0185.62.2.139

Simpson, E .H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B (Methodological) 13*, 238–241.

Statista (2015). *Anzahl der Ausländer in Deutschland nach Herkunftsland 2014 (Stand: 31. Dezember 2014).* Retrieved from http://de.statista.com/statistik/daten/studie/1221/umfrage/anzahl-der-auslaender-in-deutschland-nach-herkunftsland

Stern, S. E., & Faber, J. E. (1997). The lost e-mail method: Milgram's lost-letter technique in the age of the Internet. *Behavior Research Methods, Instruments, and Computers, 2*, 260–263. doi:10.3758/BF03204823

Tebbe, E. A., Moradi, B., & Ege, E. (2014). Revised and abbreviated forms of the Genderism and Transphobia Scale: Tools for assessing anti-trans prejudice. *Journal of Counseling Psychology, 61*, 581–592. doi:10.1037/ cou0000043

Titzmann, P. F., Silbereisen, R. K., Mesch, G. S., & Schmitt-Rodermund, E. (2011). Migration-specific hassles among adolescent immigrants from the former Soviet Union in Germany and Israel. *Journal of Cross-Cultural Psychology, 42*, 777–794. doi: 10.1177/0022022110362756

Tykocinski, O. E., & Bareket-Bojmel, L. (2009). The Lost E-Mail Technique: Use of an implicit measure to assess discriminatory attitudes toward two minority groups in Israel. *Journal of Applied Social Psychology, 39*, 62–81. doi:10.1111/j.1559-1816.2008.00429.x

Universität Heidelberg (2011). *Studierendenstatistik Wintersemester 2011/2012* [Student statistics - Winter term 2011/2012]. Heidelberg, Germany: Ruprecht-Karls-Universität Heidelberg. Retrieved from http://www. uni-heidelberg.de/md/studium/download/ws1112_www01.pdf

Vaes, J., Castelli, L., Paladino, M.-P., Leyens, J.-P., & Giovannazi, A. (2003). On the behavioral consequences of infrahumanization: The implicit role of uniquely human emotions in intergroup relations. *Journal of Personality and Social Psychology, 85*, 1016–1034. doi:10.1037/0022-3514.85.6.1016

Vaes, J., Paladino, M.-P., & Leyens, J.-P. (2002). The lost e-mail: Prosocial reactions induced by uniquely human emotions. *British Journal of Social Psychology, 41,* 521–534. doi:10.1348/014466602321149867

Wagner, U., & Machleit, U. (1986). 'Gastarbeiter' in the Federal Republic of Germany: Contact between Germans and migrant populations. In M. Hewstone & R. Brown (Eds.), *Contact and conflict in intergroup encounters* (pp. 59–78). Cambridge, MA: Basil Blackwell.

Webb, E., Campbell, D., Schwartz, R. & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago, IL: Rand McNally.

Webb, T. L. (2011). Advice-taking as an unobtrusive measure of prejudice. *Behavior Research Methods, 43*, 953–963. doi:10.3758/s13428-011-0122-8

Weigel, R. H., & Newman, L. S. (1976). Increase Attitude-Behavior Correspondence by Broadening the Scope of Behavioral Measure. *Journal of Personality and Social Psychology, 33*, 793–802. doi:10.1037/0022-3514.33.6.793

Worbs, S., Bund, E., Kohls, M., & Babka von Gostomski, C. (2013). *(Spät-)Aussiedler in Deutschland: Eine Analyse aktueller Daten und Forschungsergebnisse.* Forschungsbericht 20. [Late repatriates in Germany: Analysis of recent data and results. Research report 20]. Nürnberg, Germany: Bundesamt für Migration und Flüchtlinge.

Appendix A

Stimulus materials: Names, e-mail addresses, and manipulation check

Table A1

List of Names and E-Mail Addresses

Nationality/ Ethnicity	Sex	Name	E-mail address
German	female	Lisa Schmidt	lisa.schmidt2@freenet.de
German	male	Peter Müller	peter.mueller93.1@web.de
American	female	Grace Connor	grace_connor@rocketmail.com
American	male	Tyler Smith	tyler.smith@freenet.de
Italian	female	Giulia Ferro	giulia.ferro@web.de
Italian	male	Guiseppe Rossi	giusepperossi93@gmail.com
Russian	female	Anastasija Iwanow	anastasija.iwanow4@gmail.com
Russian	male	Wladimir Smirnow	wladimir.smirnow@gmx.de
Turkish	female	Ayse Yilmaz	ayse_yilmaz1@gmx.de
Turkish	male	Ahmet Atatürk	ahmetatatuerk@yahoo.de

Note. One set of names, randomly coupled with e-mail providers, was used. First names were common among 18- to 20-year olds.

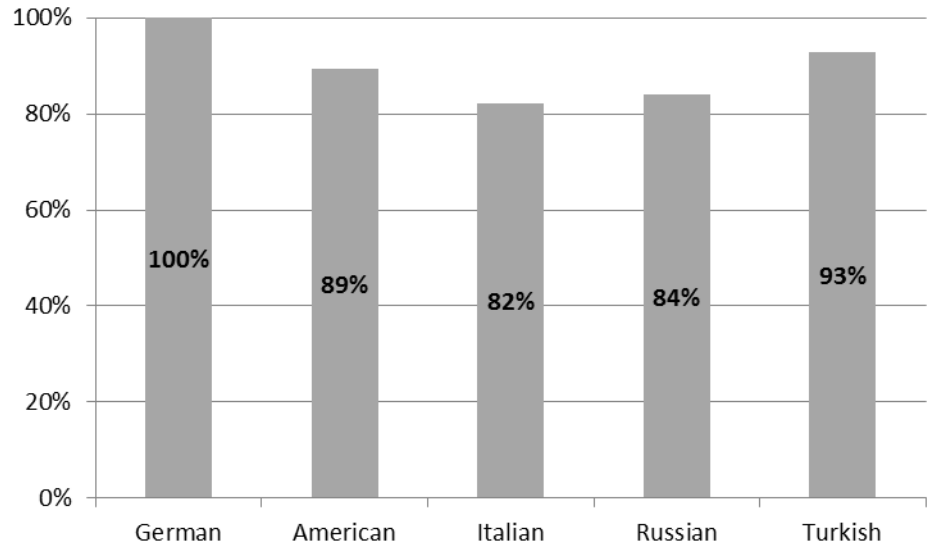

Figure A1. Stimulus manipulation check from 28 students who supplied open-ended responses to the ethnicity/nationality question. Correct identification rates were computed across two names per ethnic group. Foreign group labels were coded as correct only if the intended group was explicitly mentioned, either by itself or in combination (e.g., "Australian/Anglo-Saxon" for American names or "Caucasian" for Russian names were counted as incorrect but "US/UK" or "Russian/Balkan" were counted as correct).

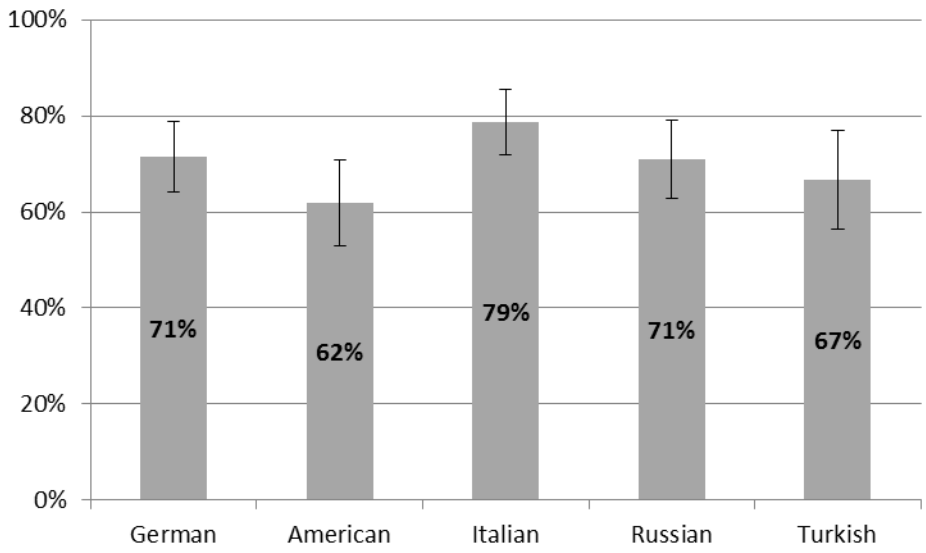

Figure A2. Average ratings of trustworthiness and 95%-confidence intervals from 28 pretest participants who provided open-format numerical ratings (as percentage points) for the trustworthiness of each e-mail address.

Appendix B
Stimulus materials: Exemplary wording of e-mails

Example e-mail body, translated from German to English. Five e-mail templates were randomly coupled with five names within participants before distribution.

Subject: Your ad / wg-gesucht.de

Text body: Dear Mr. [name of lessor] / Dear Ms. [name of lessor]

I am interested in your apartment listing of [date], posted on wg-gesucht.de. I am an 18-year-old student and currently live with my parents, who would provide a monetary bond for the apartment. Is it possible for me to inspect the apartment?

Kind regards,
[name of purported tenant]

Appendix C
Stimulus materials: Original German e-mail templates

Template #1:

Sehr geehrte/r Herr/Frau …!
Ich interessiere mich für Ihr Angebot vom … auf wg-gesucht.de. Ich bin Student/-in, 18 Jahre alt und lebe zurzeit noch bei meinen Eltern, die auch für die Wohnung bürgen würden. Kann ich zu einem Besichtigungstermin kommen?
Mit freundlichen Grüßen, …

Template #2:

Sehr geehrte/r Herr/Frau …,
Bestünde die Möglichkeit die Wohnung, die Sie auf wg-gesucht.de inseriert haben zu besichtigen? Ich bin 20 Jahre alt und studiere seit kurzem in Heidelberg. Falls Sie einen Bürgen für die Wohnung benötigen, würden sich meine Eltern dazu zur Verfügung stellen.
Herzliche Grüße, …

Template #3:

Hallo Frau/Herr…!
Ich habe Ihre Anzeige auf wg-gesucht.de gelesen und bin sehr an der Wohnung interessiert. Könnte ich zur Besichtigung der Wohnung kommen? Zu meiner Person: Ich bin ein/e 20-jährige/r Student/in und komme aus Hamburg. Falls es nötig sein sollte, wären meine Eltern bereit, die Bürgschaft für die Wohnung zu übernehmen.
Mit freundlichen Grüßen, …

Template #4:

Liebe/r Frau/Herr …,
Könnte ich die Wohnung (Anzeige bei wg-gesucht) besichtigen? Ich studiere in Heidelberg und bin 18 Jahre alt. Wegen der Miete besteht kein Grund zur Sorge, da meine Eltern für die Wohnung aufkommen würden.
Mit freundlichen Grüßen, …

Template #5:

Sehr geehrte/r Herr/Frau…!
Ich bin vor kurzem auf Ihre Anzeige bei wg-gesucht gestoßen und hätte großes Interesse bald zu einer Besichtigung zu kommen. Ich bin 19 Jahre alt und wohne derzeitig noch bei meinen Eltern in Hamburg. In Kürze nehme ich mein Studium in Heidelberg auf. Meine Eltern ließen sich überzeugen eine Bürgschaft zu übernehmen. Ich würde mich sehr über einen Besichtigungstermin freuen.
Mit freundlichem Gruß, …