NBER WORKING PAPER SERIES

INFREQUENT IDENTITY SIGNALS AND DETECTION RISKS IN AUDIT CORRESPONDENCE STUDIES

Catherine Balfe
Patrick Button
Mary Penn
David Schwegman

Working Paper 28718
http://www.nber.org/papers/w28718

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2021

Patrick Button thanks the National Institutes of Health for funding via a postdoctoral training grant to the RAND Corporation (5T32AG000244-23), which funded their research from 2018 to 2019. We thank seminar participants at the APPAM, AEA, and AREUEA conferences, and Tulane University, and we thank Lei Gao, Andrew Hanson, Joanna Lahey, and Mike Martell for helpful comments, guidance, and discussions. We also thank Eva Dils and Ilan Gressel for excellent research assistance. This study was approved by Syracuse University's IRB (#18-176.) The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Infrequent Identity Signals and Detection Risks in Audit Correspondence Studies
Catherine Balfe, Patrick Button, Mary Penn, and David Schwegman
NBER Working Paper No. 28718
April 2021
JEL No. C93,J15,J16,J71,K42,Z13

ABSTRACT

Audit correspondence studies are field experiments that test for discriminatory behavior in active markets. Researchers measure discrimination by comparing how responsive individuals ("audited units") are to correspondences from different types of people. This paper elaborates on the tradeoffs researchers face between sending audited units only one correspondence and sending them multiple correspondences, especially when including less common identity signals in the correspondences. We argue that when researchers use audit correspondence studies to measure discrimination against individuals that infrequently interact with audited units, they raise the risk that these audited units become aware they are being studied or otherwise act differently. We present the result of an audit correspondence study that demonstrates how this detection can occur when researchers send more than one correspondence from an uncommon minority group. We show how this detection leads to attenuated (downwardly biased) estimates of discrimination.

Catherine Balfe
Education Research Alliance
1555 Poydras Street
7th Floor, Room 701
New Orleans, LA 70112
cbalfe@tulane.edu

Patrick Button
Department of Economics
Tulane University
6823 St. Charles Avenue
New Orleans, LA 70118 and
NBER
pbutton@tulane.edu

Mary Penn
Department of Economics
Tulane University
6823 St. Charles Avenue
Tilton Hall 206
New Orleans, LA 70118
mpenn2@tulane.edu

David Schwegman
Department of Public Administration and Policy
School of Public Affairs
American University
4400 Massachusetts Avenue NW
Washington, DC 20016
schwegma@american.edu

Audit correspondence studies are a popular type of field experiment used to detect and measure discrimination. Audit studies are considered the gold standard for detecting discrimination, which otherwise may go undetected in more traditional methods, such as surveys or in-person interviews, because such behavior is often illegal or not revealed due to social desirability bias.

In audit correspondence studies, the researcher contacts, most frequently using email, individuals (e.g., politicians, landlords) or businesses to inquire about a service, seek employment, or ask a question. These individuals or businesses are the "audited unit." In these correspondences, the researcher randomly assigns signals, such as names or other self-disclosed characteristics, that signal a specific identity (e.g., race, ethnicity, gender) or attribute (e.g., education level, occupation) of the inquirer. Suppose the audited units respond less to a correspondence containing one type of signal (e.g., African American) compared to correspondences containing a different signal (e.g., white). In that case, this provides evidence that the audited units are discriminating against individuals possessing this signal. In a famous example, economists Bertrand and Mullainathan (2004) sent resumes containing stereotypical white names (e.g., Emily, Greg) and stereotypical black names (e.g., Lakisha, Jamal) to job openings. They found that resumes containing white names received 50 percent more callbacks for interviews.

In recent years, there has been a proliferation in the use of audit studies to study discrimination in new and emerging markets (e.g., the "Sharing Economy"), as well as discrimination against relatively smaller minority groups (e.g., LGBTQ+ people) and based on multiple treatments such as intersectional identities (Rooth 2021; Baert 2018; Edelman et al. 2017). In these new contexts, the proper design of audit correspondence studies is even more critical. Sending correspondences that (1) signal that the inquirer belongs to a community that are

infrequent customers, constituents, or clients of the audited unit (e.g., same-gender married couples), or (2) contain specific attributes that are unlikely to be disclosed (e.g., unsolicited information) increases the risk of detection by audited units.

Detection can lead to "spillovers," where audited units do not treat the different correspondences they receive independently. For example, an audited unit that gets an unusual email may treat future emails as suspicious. This detection leads to biased estimates of discrimination: audited units may change their behavior if they think they are being monitored (i.e., an observer effect) or think the correspondence they get is unusual. We argue that detection risk is much higher when using an audit field experiment to study discrimination against groups infrequently encountered by audited units, particularly if researchers send more than one piece of similar correspondence from those groups. We also argue that detection risk is higher when more correspondences are sent to each audited unit, especially if the time between correspondence pieces is short, and the correspondences are similar.

To show how exposing audited units to multiple infrequent identity signals and multiple correspondence pieces raises detection risk, we present the results of an email-based audit correspondence study of sexual orientation discrimination by mortgage loan originators in the United States. We find that being exposed to more than one infrequent identity signal (two same-gender couples inquiring about mortgage loans) significantly increases the likelihood of detection. We also find evidence that sending more than two pieces of correspondence in total could also raise detection risk. Audited units exposed to either more than one email by same-gender married couples or three or more correspondence pieces behave differently, biasing discrimination estimates towards zero. When we restrict our sample to only consider when the audited unit has been exposed to a at most two correspondence pieces and only one email from a same-gender

married couple, then the discrimination estimate is two to three times greater relative to the sample where detection occurred. Thus, the bias caused by detection in our case study was severe.

Our finding that detection can easily occur and leads to significant bias is critical for researchers to consider when designing correspondence studies. Researchers face tradeoffs when deciding how many pieces of correspondence to send to each audited unit. Sending more correspondence pieces to each audited unit increases statistical power and allows for within-unit testing but increases detection risk (and thus bias). Detection risk is often unclear, making it difficult for researchers to select the best number of correspondence pieces to send to each audited unit. We argue that this detection risk and resulting bias could be significant when researchers send two or more pieces of correspondence from groups that audited units infrequently interact with or are otherwise atypical. We argue that detection risk also exists when researchers send more than two pieces of correspondence, especially in a short period of time and when the correspondence is otherwise similar.

How Many Pieces of Correspondence Should Researchers Send?

Researchers face tradeoffs in how many pieces of correspondence they send to each audited unit. There are several benefits to sending more correspondences, including increased statistical power, reduced data collection costs, and the ability to make within-audited-unit comparisons. However, sending more than one correspondence increases the risks of detection or spillovers, both of which could bias estimates of discrimination and invalidate the study. It also imposes additional time costs or other costs onto audited units. Below, we summarize these costs and benefits and existing research on them. This paper's contribution is to illuminate the risks and costs of detection, a concern for which little is known.

Statistical Power

The primary benefit of sending more pieces of correspondence is that it can significantly increase statistical power. Sending, for example, two pieces per audited unit rather than one will double the sample size. Importantly, however, this will not double power. The two pieces of correspondence within the same audited unit are not statistically independent.[1] Nevertheless, increasing power by, say, 80% by sending twice as many pieces of correspondence to each audited unit is a substantial benefit.

Sufficient statistical power is required to detect meaningful levels of discrimination. Including multiple correspondence per audited sample can be critical in certain study contexts where there is a finite sample of audited units (Vuolo, Uggen, and Lageson 2018), or experiments that include multiple groups or treatment arms. Adequate power is even more crucial for testing moderators of discrimination or focusing on subgroups, such as detecting intersectional

[1] See Lahey and Beasley (2018) for discussion and formulas relating to the inter-correlation between clusters.

discrimination. Sending additional correspondence per unit could also be necessary to meet the required sample size if few audited units are available.

Reduced Data Collection Costs

Sending multiple pieces of correspondence to each audited unit often decreases data collection costs per observation. The fixed costs of collecting information on an audited unit (e.g., locating a job ad, finding an email address) do not occur again when sending additional pieces of correspondence to each unit. Sending more correspondence to each audited unit may mean that the researcher requires fewer audited units in their study, reducing these fixed costs.

Within-Audited-Unit Comparisons

Sending multiple pieces of correspondence to each audited unit allows more within-audited-unit comparisons. If researchers only send one piece of correspondence, there is no way to compare responses within one audited unit. Sending multiple correspondences per audited unit enables the researcher to include audited unit fixed effects, which control for differences between audited units. Thus, the researcher can control for important factors such as organizational culture, community demographics, local laws, and audited unit financial resources. Importantly, most studies do not require the inclusion of these fixed effects since the correspondence sent to audited units is randomized. However, if it were not, this would allow the researcher to control for fixed differences between audited units that may differ based on treatment status. Regardless, it is desirable to test the robustness of results to audited unit fixed effects.

Spillovers Within an Audited Unit

Sending additional correspondence can lead to spillovers, where pieces of correspondence within the same audited unit affect each other (“spillovers”), which violates a required assumption

for an experiment to have an unbiased estimated treatment effect: the Stable Unit Treatment Value Assumption (SUTVA). Simply put, SUTVA means that each piece of correspondence is unaffected by other pieces of correspondence.

Phillips (2019) best demonstrates a common possible violation of SUTVA in audit correspondence studies that send multiple pieces of correspondence, in this case resumes to job openings. Spillovers can occur in this case for two reasons. First, resumes could be directly compared, rather than being evaluated independently of each other. Second, businesses may decide to hire more applicants, which may occur since the applicant pool appears stronger due to the additional resumes sent by the researcher. Phillips (2019) shows that some prior resume correspondence studies that sent four resumes to each job opening, such as Bertrand and Mullainathan (2004), had spillover issues, violating SUTVA and understating discrimination estimates by about 20% on average.

Detection Risks

A common fear is that the experiment is detected, i.e., audited units think something odd is happening such as some enforcement action, a bot sending emails, or even believing that they are in an experiment. If the audited units detect the experiment in some way, then they could alter their behavior. *A priori*, it is unclear how audited units would react if they knew the inquiries were fake, someone (or something) was monitoring their behavior, or they simply thought something was unusual. They could act in a way that they think is socially desirable, such as responding to all correspondence. Alternatively, they could instead see all pieces of correspondence as unusual and ignore them. Thus, changes in behavior can be thought of as statistical noise, which will downwardly bias estimates of discrimination (i.e., attenuation bias).

While detection risks are higher if the correspondence is unusual, which is why researchers typically put much effort into making their correspondence seem like what the audited unit gets typically, there is still detection risk in studies that use normal-looking correspondence if the researcher includes signals that the inquirer (or applicant) belongs to a group with whom the audited unit rarely interacts. Consider the audited unit from our audit study presented in this article: mortgage loan originators (MLOs), who help customers get mortgage loans. It may be relatively unlikely, but not unusual, for an MLO to interact with same-gender married couples, given their less frequent incidence in the population. However, it may be improbable for an MLO to receive multiple emails in a short period from multiple infrequent customers (e.g., several same-gender couples). It could also be unusual to receive emails from frequent (e.g., different-gender couples) and infrequent customers (e.g., same-gender couples) that contain the same attributes.

Even if the researcher sends normal-looking correspondence containing signals that the applicant is a member of the audited unit's normal customer base, it may be odd to receive *multiple* very similar correspondences (containing similar language or possessing a similar structure) too closely together. Thus, if a researcher decides that they would like to—or, given a finite sample of audited units, need to—send multiple pieces of correspondence, they next need to decide how much to vary the syntax and structure of each correspondence, as well as the time between when they send the correspondences (order should be randomized). The researcher should not vary the correspondences in a way that introduces unintended signals into the correspondences, e.g., unintentionally varying the formality of the correspondences. Such changes can undermine the study design.

In the following sections, we detail an audit correspondence study that faced this detection issue. The detection issues were severe in this study—the estimated measure of discrimination was

likely less than half of what is in a sample that is likely unaffected by detection. Our results demonstrate that detection issues can severely undermine the internal validity of a study, so, it may not be advisable to send too many pieces of correspondence as the costs outweigh the benefits.

Experimental Design

To test the likelihood of detection and spillover effects from sending multiple emails signaling infrequent identities, we used data from our audit correspondence study on sexual orientation discrimination in access to mortgage loans. Specifically, we conducted an email correspondence study testing if mortgage loan originators discriminate against same-gender couples requesting assistance obtaining a mortgage loan. We initially designed our experiment as a pilot study. However, it serves the additional, unexpected purpose of being a case study in how the number and type of correspondences sent can cause detection and spillover issues.

We conducted our small-scale study with a convenience sample of 118 MLOs across all regions of the United States. To generate our sample, we collected publicly posted MLO email addresses through various websites (individual bank websites, Yellow Pages, Better Business Bureau, LinkedIn).

In our experiment, each MLO received four emails, one week apart, in random order: one email from a same-gender male couple, one from a same-gender female couple, one from a different-gender couple with a male email sender, and one from a different-gender couple with a female email sender. We conducted our experiment between September and October 2018.

Our experiment's general design follows similar email-correspondence studies (Ahmed and Hammarstedt 2009; Hanson et al. 2016; Schwegman 2019). In each email, following Ahmed and Hammarstedt (2009), we signaled sexual orientation by having each fictitious applicant introduce

themselves as well as their wife or husband. We did not vary race or ethnicity.[2] We randomly included three attribute signals correlated with the perceived risk of loan applicants: credit scores, employment status, and children or fertility signals. Our Online Appendix presents the templates for our emails and the email components and additional details about the experimental design.

Our primary measure of discrimination is whether the MLO responded to the original email. We considered a response to be one that appears to be written by a human. Thus, out-of-office or other automated responses were not counted as a response. We only consider responses sent within two weeks of the original inquiry email (following Hanson et al. 2016).

In this experiment, we consider same-gender customers to be infrequent customers. Since it is not unrealistic for an MLO to interact with a same-gender couple, a single inquiry from a same-gender couple is unlikely to raise concern. However, sending multiple somewhat similar correspondences and, most notably, multiple similar correspondences from same-gender couples, especially in a short period, could increase detection risks. Unless the MLO operates in specific urban markets with a large population of same-gender couples, it is unlikely that an MLO will receive an inquiry from multiple same-gender couples in the same month.

Statistical Estimation Methodology

Before we present how we determine how detection affects our estimates of discrimination, we first present how we estimate discrimination in our experiment. Our primary outcome is if there was a non-automated response. To quantify differences in response rates by couple type, we use a linear regression model of the form:

[2] Thus, we use common first and last names that generally signal that the individual is likely to be white. We used the first names from Friedman et al. (2013), which used the 20 most popular names in the United States in the 1980s. The list only includes names that strongly signal gender. We used common last names from Neumark, Burn, and Button (2019). See the Online Appendix for a complete list of names.

$$Response_i = \beta_0 + \beta_1 MM_i + \beta_2 FF_i + \beta_3 FM_i + Controls_i \beta_4 + \varepsilon_i \quad [1]$$

Where *i* indexes each email, MM_i *(*FF_i*)* is an indicator variable for same-gender male (female) spouses, and FM_i is an indicator variable for different-gender spouses where the wife sent the email, with the excluded category being different-gender spouses where the husband sends the email. *Controls* is a vector of control variables, including randomized credit scores for the couple and fixed effects for different randomized email components and email order. We also estimate a regression like [1] but pooling all same-gender couples together rather than splitting by gender.

To test how detection could have impacted our estimates of discrimination (the regression coefficients β_1 and β_2), we run our regressions using the three samples of the data described in Table 1: the full sample (Panel A), restricted sample #1 (Panel B), and restricted sample #2 (Panel C). Each panel in Table 1 shows eight possible email orders based on sending two same-gender ("same") and two different-gender ("different") emails.

[Table 1 about here.]

To test if MLOs change their behavior when receiving (and after receiving) a second same-gender email, we compare the full sample (Panel A) to restricted sample #1 (Panel B). In restricted sample #1, we drop all second same-gender emails and any emails we sent after the second same-gender email ("DROPPED"). Comparing the results between the full sample and restricted sample #1 allows us to test if discrimination estimates differ if the sample includes cases where the audited unit (MLO) received more than one piece of correspondence from an uncommon group, i.e., same-gender married couples.

We also consider restricted sample #2 (Panel C), which restricts the sample further by removing any round 3 emails not already excluded from restricted sample #1. Restricted sample #2 only includes the first two rounds. Comparing discrimination estimates between restricted

samples #1 and #2 allows us to determine if sending a third email, conditional on only sending one email total from a same-gender married couple, causes any detection issues compared to just sending up to two emails.

Results

Before presenting more formal estimates of how our discrimination estimates may have been affected by detection and spillovers, we first present response rates by round. Figure 1 presents the response rate to emails from round 1 (the first email the MLO received) to round 4 (the fourth email the MLO received, four weeks later). Figure 1 shows that many MLOs changed their behavior between rounds and did so differently if the email was from a same-gender or different-gender couple. The MLOs' response rate to different-gender couples was relatively constant in the first three rounds: 70.8%, 73.7%, and 72.1%. However, for same-gender couples, the response rate gradually increased from 42.9% to 50.8% to 60.4%. In round 4, response rates changed dramatically. The response rate for same-gender couples dropped to 42.9%. The response rate for different-gender couples dropped far below this, to 26.2%, the lowest response rate for either group for any round. Thus, discrimination appears to decrease rapidly by round, to the point that same-gender couples appear preferred in round 4.

[Figure 1 about here.]

Table 2 presents our regression results using two regressions, one with same-gender female and male couples combined in one indicator variable (odd columns) and one with same-gender female and male couples analyzed separately (even columns). Columns 1 and 2 use the full sample, columns 3 and 4 use restricted sample #1, and columns 5 and 6 use restricted sample #2.

[Table 2 about here.]

Starting with the full sample and Column 1, we find that same-gender couples were 11.4 percentage points less likely to receive a response compared to different-gender couples (significant at the 1% level). In Column 2, we separate the same-gender indicator into separate indicator variables for same-gender male couples and same-gender female couples. While we find large negative coefficients, these differences are not statistically significantly different from zero.

When we restrict the sample so that MLOs have only received one same-gender email (restricted sample #1, Columns 3 and 4), we generally find a larger estimate of discrimination, although the estimates are noisy. The estimated level of discrimination faced by same-gender couples (Column 3) slightly increased to 15.6 percentage points. That is, same-gender couples are 15.6 percentage points less likely to receive a response. However, this estimate has a large standard error, and it is therefore not statistically significant. For male same-gender couples, the estimate is a massive 43.8 percentage point lower response rate (significant at the 1% level), while for same-gender female couples, the response rate is the same as for different-gender couples.

We then further restrict our sample to restricted sample #2 (Columns 5 and 6), which drops any remaining round 3 emails not already dropped from restricted sample #1. The results again increase in magnitude, showing even more discrimination, and the estimates show discrimination more consistently across regressions and by gender. Column 5 shows that the response rate was 27.2 percentage points lower for same-gender couples in general (significant at the 1% level), with discrimination estimates being similar by gender (Column 6).

The reason for this change in the estimated level of discrimination faced by same-gender couples across our samples is clear from Figure 1. Restricted sample #1 (#2) drops part (all) of round 3 and all of round 4. These rounds had wildly different response rates and response rate differences by sexual orientation, especially for round 4.

Discussion and Conclusion

The results from our audit correspondence study show that some MLOs became aware of our study because (a) they received multiple similar correspondences in a short period of time and (b) some of these correspondences included inquiries from infrequent customers (i.e., same-gender couples). MLOs changed their behavior in two ways: responding to same-gender couples more often (social desirability bias) and, by email round 4, responding less often to everyone. Both these behaviors led to biased estimates of discrimination. Our experiment shows that once we remove observations tainted by possible detection (we use restricted samples #1 and #2), estimated discrimination increases by around two to three times.

Our results suggest that researchers should reduce an audited unit's exposure to multiple infrequent signals to minimize the detection risks and resulting spillovers and bias. We suggest sending either one piece of correspondence, randomized between treatment versus control, or sending one treatment and one control piece of correspondence. While it may be tempting to send additional pieces of correspondence to each audited unit to increase the sample size and improve scarce statistical power, the risk of bias from detection or spillovers could be very high. Sending three or four correspondence pieces could perhaps avoid these high costs of detection or spillovers only if all the following apply: the audited units frequently get similar correspondence, the correspondence is distinct from each other, there are sufficient time delays between correspondence, and the spillover concerns detailed in Phillips (2019) do not apply.

References

Ahmed, Ali M., and Mats Hammarstedt. 2009. "Detecting Discrimination against Homosexuals: Evidence from a Field Experiment on the Internet." *Economica* 76 (303): 588–97. https://doi.org/10.1111/j.1468-0335.2008.00692.x.

Baert, Stijn. 2018. "Hiring Discrimination: An Overview of (Almost) All Correspondence Experiment Since 2005." In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. Michael Gaddis, 63-81. New York: Springer.

Bertrand, Marianne, and Esther Duflo. 2017. "Field Experiments on Discrimination." In *Handbook of Economic Field Experiments*, edited by Abhijit Vinayak Banerjee and Esther Duflo, 309–93. Elsevier. https://doi.org/10.1017/CBO9781107415324.004.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94(4), 991–1013. https://doi.org/10.1257/0002828042002561.

Button, Patrick, Eva Dils, Benjamin Harrell, Luca Fumarco, and David J. Schwegman. 2021. "Gender Identity, Race, and Ethnicity Discrimination in Access to Mental Health Care: Preliminary Evidence from a Multi-Wave Audit Field Experiment." *National Bureau of Economic Research Working Paper,* No. 28164.

Edelman, Benjamin, Michael Luca, and Dan Svirsky. 2017. "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment." *American Economic Journal: Applied Economics* 9(2): 1-22.

Friedman, S., Reynolds, A., Susan Scovill, F. R., Brassier, R. C., & Ballou, M. (2013). *An estimate of housing discrimination against same-sex couples.* Office of Policy Development and Research, United States Department of Housing and Urban Development (HUD). Retrieved from https://www.huduser.gov/portal/Publications/pdf/Hsg_Disc_against_SameSexCpls_v3.pdf

Gaddis, S. Michael. 2017. "How black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies. " *Sociological Sciences* (4): 469-489.

Hanson, Andrew, Zackary Hawley, Hal Martin, and Bo Liu. 2016. "Discrimination in Mortgage Lending: Evidence from a Correspondence Experiment." *Journal of Urban Economics* 92: 48–65. https://doi.org/10.1016/j.jue.2015.12.004.

Lahey, Joanna N., and Ryan Beasley. 2018. "Technical Aspects of Correspondence Studies." In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. Michael Gaddis, 81–101. New York: Springer.

Neumark, David, Ian Burn, and Patrick Button. 2019. "Is It Harder for Older Workers to Find Jobs? New and Improved Evidence from a Field Experiment." *Journal of Political Economy*

127 (2): 922–70. https://doi.org/10.1086/701029.

Pager. Devah. 2016. "Are firms that discriminate more likely to go out of business?" *Sociological Sciences* 3: 849-859.

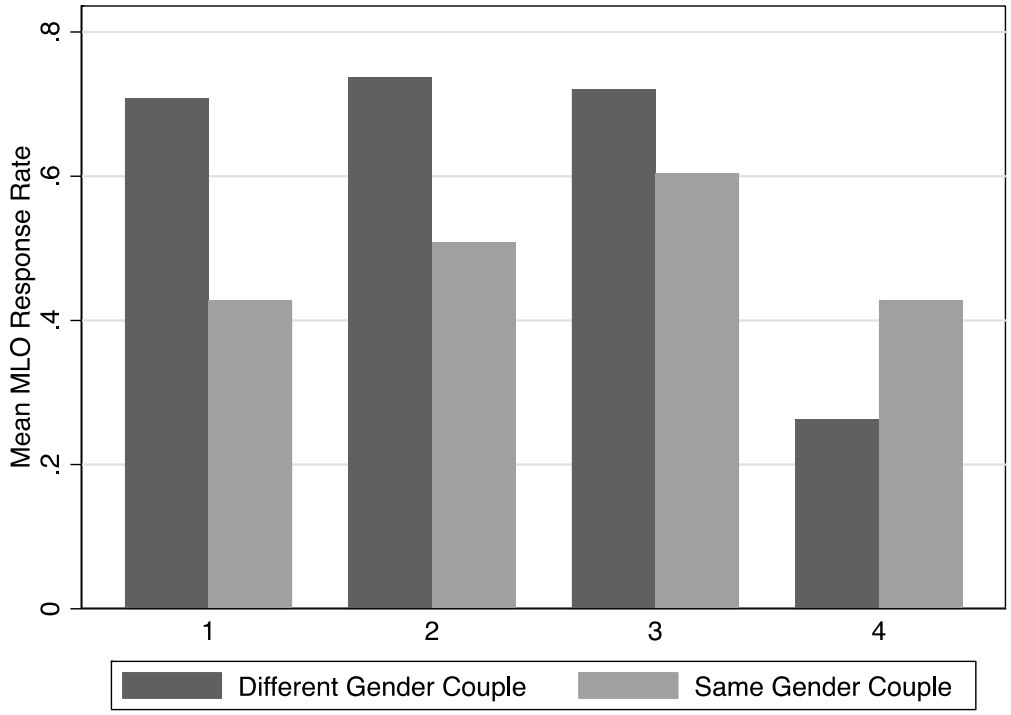
Phillips, David C. (2019). "Do Comparisons of Fictional Applicants Measure Discrimination When Search Externalities Are Present? Evidence from Existing Experiments. " *Economic Journal* 129(621): 2240-64. https://doi.org/10.1111/ecoj.12628.

Rooth, Dan-Olof. 2021. "Correspondence testing studies: What is there to learn about discrimination in Hiring." *IZA World of Labor*. DOI: 10.15185/izawol.58.v2

Schwegman, David. 2019. "Rental Market Discrimination Against Same-Sex Couples: Evidence From a Pairwise-Matched Email Correspondence Test." *Housing Policy Debate* 29 (2): 250–72. https://doi.org/10.1080/10511482.2018.1512005.

Vuolo, Mike, Christopher Uggen, and Sarah Lageson. 2018. "To Match or Not to Match? Statistical and Substantial Consideration in Audit Design and Analysis." In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. Michael Gaddis, 119-142. New York: Springer.

Figure 1: Response Rates by Email Round and Sexual Orientation, Full Sample



Notes: Each round is about one week apart. Round 1 was at 1 pm CST on November 18, 2018; round 2 was at 2 pm CST on November 26, 2018; round 3 was at 11 am CST on December 2, 2018; round 4 was at 9 am CST on December 10, 2018.

Table 1: Comparison of Full and Restricted Samples

Panel A: Full Sample

Email Order	Round 1	Round 2	Round 3	Round 4
1	Different	Different	Same	Same
2	Different	Same	Different	Same
3	Different	Same	Same	Different
4	Same	Different	Same	Different
5	Same	Different	Different	Same
6	Same	Same	Different	Different

Panel B: Restricted Sample #1

Email Order	Round 1	Round 2	Round 3	Round 4
1	Different	Different	Same	DROPPED
2	Different	Same	Different	DROPPED
3	Different	Same	DROPPED	DROPPED
4	Same	Different	DROPPED	DROPPED
5	Same	Different	Different	DROPPED
6	Same	DROPPED	DROPPED	DROPPED

Panel C: Restricted Sample #2

Email Order	Round 1	Round 2	Round 3	Round 4
1	Different	Different	DROPPED	DROPPED
2	Different	Same	DROPPED	DROPPED
3	Different	Same	DROPPED	DROPPED
4	Same	Different	DROPPED	DROPPED
5	Same	Different	DROPPED	DROPPED
6	Same	DROPPED	DROPPED	DROPPED

Notes: See the notes to Figure 1. Different = Different-gender couple; Same = Same-gender couple; DROPPED = Not included in the sample. The sample size for the full sample is 469, 258 for Restricted Sample #1, and 204 for Restricted Sample #2.

Table 2: Regression Estimates for MLO Response, by Type of Married Couple

	(1)	(2)	(3)	(4)	(5)	(6)
Same-Gender (both)	-0.114***	…	-0.156	…	-0.272***	…
	(0.042)		(0.124)		(0.067)	
Same-Gender Male	…	-0.086	…	-0.438***	…	-0.312***
		(0.070)		(0.157)		(0.096)
Same-Gender Female	…	-0.111	…	-0.018	…	-0.291**
		(0.074)		(0.151)		(0.115)
Different-Gender,	…	0.029	…	-0.159	…	-0.061
Female Emailer		(0.086)		(0.114)		(0.124)
Sample	Full Sample		Restricted Sample #1		Restricted Sample #2	
N	469		258		204	

Notes: The response rate in the full sample for the excluded group in column (2) (different gender, male email sender) is 59.1%. We cluster our standard errors by MLO. Significantly different from zero at 1-percent level (***), 5-percent level (**), or 10-percent level (*). All regressions include fixed effects for the round and for the email template, indicator variables for if either spouse has each occupation, fixed effects for occupational tenure, and fixed effects for the number of children or if the couple is expecting a child. See the Online Appendix for more details on the experimental design and these control variables.

Online Appendix: Additional Details on the Experimental Design

Online Appendix Figure A1 presents the structure of our emails to MLOs and Online Appendix Figure A2 shows the randomized parts of our emails. We sent each MLO four emails from four different couples: a same-gender male couple, a same-gender female couple, a different-gender couple with a male email originator, and a different-gender couple female email originator. We sent emails about one week apart.

We signaled sexual orientation by having each fictitious applicant introduce themselves as well as their wife or husband. We did not vary race or ethnicity and therefore used common first and last names that generally signal that the individual is likely to be white. We included first names from Friedman et al. (2013), which used the 20 most popular names in the United States in the 1980s. The list only includes names that strongly signal gender. We use common last names from Neumark, Burn, and Button (2019). See Online Appendix Figure A3 for a complete list of names. Additionally, we randomly included three attribute signals correlated with the perceived risk of loan applicants: credit scores, employment status, and children or fertility signals.

Following Hanson et al. (2016), we signaled for creditworthiness by randomly assigning both individuals either low credit scores (mean 643, drawn from a uniform distribution between 631 to 655) and high credit scores (mean 753, range 731 to 774), each with a 50% probability. Please see Hanson et al. (2016) for a discussion on the importance of including credit score.

We randomly included occupation and tenure of employment signals as additional creditworthiness signals. Occupation and tenure of employment are both seen as important in the loan decision. We randomly assigned one of the following occupations, without replacement, for each spouse: childcare provider ($23,760), retail worker ($27,459), administrative assistant ($35,850), construction worker ($48,161), high school drama or math teacher ($52,036), registered nurse ($69,713), human resource manager ($123,510), healthcare administrator ($131,310), dermatologist ($211,390), and psychiatrist ($216,090) (U.S. Bureau of Labor Statistics, 2017).

For half of the couples, we signaled that they have one child, two children, or are expecting a child. We randomized over one child, two children, or expecting with equal probability for couples assigned a children or fertility signal, except for same-gender male couples, we randomized over having a child or having two children. We included these children or fertility signals since they may affect creditworthiness and perceptions of fertility could differ between different-gender couples, same-gender female couples, and same-gender male couples.

Online Appendix Figure A1: Email Template

1) [EMAIL SUBJECT LINE]

2) [GREETING],

My name is [MALE NAME or FEMALE NAME] and my [HUSBAND, MALE NAME or WIFE, FEMALE NAME] and I are interested in 3) [PRODUCT]. We 4) [HOME SEARCH] 5) [X] bedrooms) [*if children, then*: because we 6a) [ARE EXPECTING] or 6b) [HAVE (X CHILDREN or KIDS)].

7) [SOURCE]

8) [CREDIT SCORE] [RANDOMLY ASSIGNED SCORE WITHIN HIGH OR LOW GROUP]. I have been 9a) [OCCUPATION] for 9b) [OCCUPATION TENURE] years and [spouse name] is a 9c) [SPOUSE OCCUPATION].

10) [PLEASANTRY]

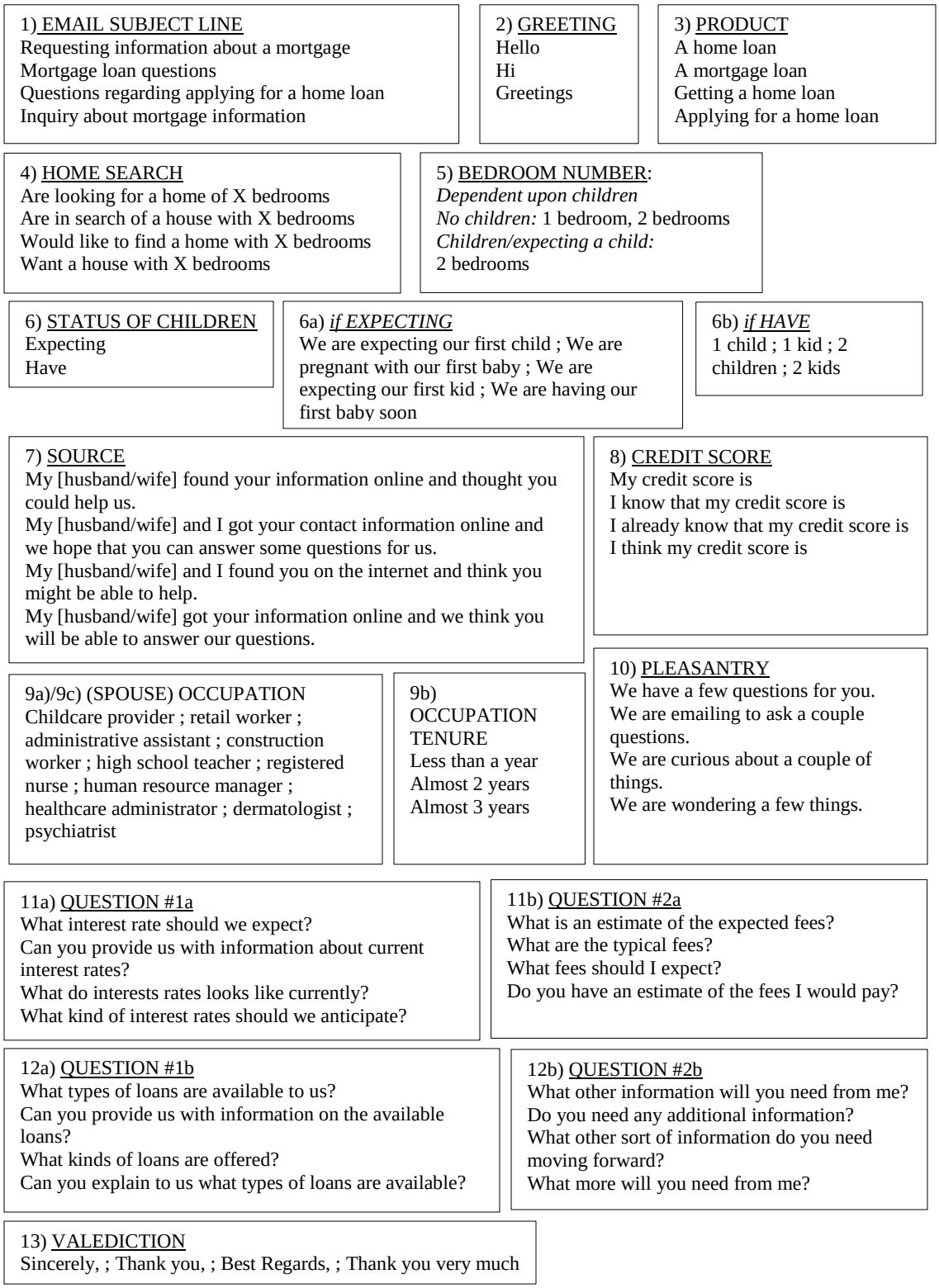
11a and 11b) [QUESTION #1a or QUESTION #2a]

12a and 12b) [QUESTION #1b or QUESTION #2b]

13) [VALEDICTION],

[NAME]

Online Appendix Figure A2: Email Components Pilot Study



Online Appendix Figure A3: Names

MALE	FEMALE	LAST
Andrew	Amanda	Adams
Brandon	Alicia	Allen
Brian	Brittany	Anderson
Christopher	Christine	Baker
David	Danielle	Campbell
Eric	Elizabeth	Clark
James	Heather	Evans
Jason	Jennifer	Hall
Jonathan	Jessica	King
Justin	Julie	Martin
Kevin	Karen	Miller
Mark	Lauren	Moore
Michael	Melissa	Nelson
Nicholas	Michelle	Phillips
Richard	Nicole	Roberts
Ryan	Rachel	Smith
Steven	Rebecca	Thompson
Thomas	Sarah	Wilson
Timothy	Stephanie	Wright
William	Tiffany	Young

Notes: First names are from Friedman et al. (2013). These are the 20 most popular girls' and boys' names in the United States from 1970 to 1985. To ensure all names strongly signal gender, we use Alicia instead of Ashley since Ashley is occasionally a male name. We use last names from Neumark, Burn, and Button (2019).