Bureaucratic Responsiveness to LGBT Americans

Kenneth Lowande∗
University of Michigan

Andrew Proctor†
Princeton University

February 27, 2019‡

Abstract
Marriage rights were extended to same-sex couples in the United States in 2015. However, anecdotes of bureaucratic non-compliance (in the form of bias or denial of license issuance) raise the possibility that de jure marriage equality has not led to equality in practice. We investigate this by conducting the only known national audit experiment of local-level marriage license granting officials in the United States. These officials vary in the constituencies they serve, as well as how they are selected, allowing us to evaluate longstanding hypotheses about bureaucratic responsiveness. Overall, we find no evidence of systematic discrimination against same-sex couples regardless of responsiveness measure, institutions, local-level ideology, or prior state legal history. We find, however, that among same-sex couples, officials tended to be more responsive to lesbian couples. Our results support recent research that has found public officials tasked primarily with service provision show less evidence of discrimination than politicians.

Word Count: 8,393

∗Assistant Professor, Department of Political Science. Contact: lowande@umich.edu
†Ph.D. Candidate, Department of Politics. Contact: aproctor@princeton.edu
‡Previous version presented at the 2019 annual meeting of the Southern Political Science Association, Austin, TX. We thank Andrew Clarke, Charles Crabtree, Thomas Gray, Hans Hassell, Adam Hughes, George Krause, Adam Levine, Scott Limbocker, Noah Nathan, Dara Strovitch, Omar Wasow, and participants at the Princeton Research in Experimental Social Science (PRESS) workshop for helpful comments and suggestions. Special thanks to Justin Fortney and Sebastian Macek for research assistance. PRESS and the Center for the Study of Democratic Politics at Princeton University provided funding. This study was approved by internal review boards at Princeton University (IRB# 10522) and the University of Michigan (IRB# 134110).
After the nationwide legalization of same-sex marriage in the U.S. in 2015, Kim Davis, a Democratic county clerk in Kentucky, refused to provide a license to a same-sex couple. This paralleled other anecdotes of discrimination, including that of then-State Supreme Court Justice Roy Moore in Alabama, who ordered county officials to cease issuing marriage licenses to same-sex couples altogether.¹ These cases raise a perennial question in democratic governance: whether the officials who must implement law will comply with directives from judges and legislators—who have limited coercive means of enforcement.

These concerns are particularly acute after watershed moments in policy change, in which bureaucrats have substantial discretion by virtue of their distance from formal principals. Executive departments failed to implement Title VI of the Civil Rights Act of 1964 for years after passage (Minta 2011). There is evidence that uneven enforcement of the Voting Rights Act of 1965 continues to influence the incorporation of under-represented minorities in elections (Marschall and Rutherford 2016). Most notably, the implementation of the Supreme Court’s Brown v. Board of Education (1954) decision required decades-long action on the part of prosecutors, activists, legislators, and even presidents (Daugherity and Bolton 2008).

Similarly, while Obergefell v. Hodges (2015) changed the status of same-sex marriages overnight in 14 states, the decision had to be implemented by over 6,000 public officials with vastly different constituencies, political principals, and little-to-no formal oversight. To investigate the responsiveness of bureaucrats to LGBT Americans, we conduct a national audit experiment of marriage license granting officials in United States. Specifically, we investigate whether street-level bureaucrats are less responsive to same-sex couples when they request information about the process of obtaining a marriage license.

Measuring differential service provision across sexual orientation has important consequences for scholarship on LGBT politics, executive accountability, and public policy. To our knowledge, this is the first systematic analysis of bureaucratic responsiveness to LGBT people in the United States and the first audit experiment of sexual orientation discrimination conducted since legalization. Thus, beyond the descriptive task of measuring potential discrimination, our study inves-

tigates whether bureaucrats comply with major policy change. In addition, since the rules governing the selection of license granting officials vary by locality, our analysis speaks to longstanding questions about whether the accountability mechanisms of public officials influence service provision. Institutional differences in responsiveness are usually explored in the context of legislative and judicial decision-makers. However, most citizen-government interactions involve executive officials, so it is important to understand how these mechanisms impact service provision. Finally, since states vary in their prior legality of same-sex marriage and other anti-LGBT laws, we assess the historical legacy of these laws on local compliance. The historical influence of geographic context on local racial attitudes (e.g. Acharya, Blackwell, and Sen 2016), for example, suggests these laws may have long-term impact on the treatment of same-sex couples.

Overall, in contrast to persistent evidence that public officials are less responsive to marginalized groups (e.g. Mendez and Grose 2018; Gell-Redman et al. 2018), we find little evidence of systematic differences across sexual orientations of unwed couples. Specifically, the 95% confidence intervals of differences in response rates, three measures of response quality, and even congratulatory language are not distinguishable from zero, are inconsistently signed, and typically do not include a magnitude effect greater than 3 percentage points. These estimates are relatively stable across estimation procedures and consistent across selection institutions and state legal history. We find some evidence that officials were less responsive to gay men relative to lesbian couples, however, this effect is largely driven by the fact that officials were systemically less responsive to male email senders, regardless of partner. Our results are consistent with emerging evidence that public officials concerned primarily with service provision exhibit less evidence of bias when dealing with citizens (e.g. Einstein and Glick 2017; Porter and Rogowski 2018; Jilke, Dooren, and Rys 2018). More broadly, our evidence does not suggest that localized anecdotes of discrimination are systematically representative.

**Same-sex Marriage and Bureaucratic Responsiveness**

Same-sex marriage in the U.S. is an ideal case for addressing proposed mechanisms for bureaucratic responsiveness. Since a large majority of citizens are married at least once, it poses a broadly-applicable administrative burden (Moynihan 2015). Classic agency models predict behavioral dif-
ferences based on selection mechanism (e.g. Besley 2006). Research on other historically marginalized groups suggests systematic differences in service provision. The historical moment raises questions about long-term compliance with judicial directives. Importantly, these differences are difficult to identify in an observational setting. A dearth of centralized data sources and under-reporting of discrimination cases necessitates an audit experiment.

This approach is common, but past work does not address the questions we raise. Much work investigates whether public officials are more or less responsive to marginalized groups, typically racial and ethnic minority constituents. In studies requesting information about voting, studies find that Black and Latino constituents receive fewer replies and less accurate information from public officials than putatively white constituents (Butler and Broockman 2011; White, Nathan, and Faller 2015). This is consistent with other work that find Blacks and Latinos face discrimination in the labor market (e.g. Bertrand and Mullainathan 2003; Pager, Bonikowski, and Western 2009), as consumers (Ayres and Siegelman 1995; Doleac and Stein 2013; Turner et al. 2013) and in higher education (Milkman, Akinola, and Chugh 2014). A notable exception to these findings is Einstein and Glick (2017), who find mixed evidence for racial discrimination among public housing officials.

These studies are informative about the nature and degree of discrimination facing racial and ethnic minorities, but they cannot speak to whether other marginalized constituencies, such as lesbian, gay, bisexual and transgender (LGBT) people, face similar forms of discrimination when interacting with public officials. The first known audit experiment examining discrimination based on sexual orientation was conducted by Barry Adam in 1981 (Badgett et al. 2008). More recent work has argued sexual minorities face discrimination in the labor market (Crow et al. 1991; Hebl et al. 2002; Weichselbaumer 2001; Bailey, Wallace, and Wright 2013). Van Hoye and Lievens (2003) provides an exception. Other audit experiments find same-sex couples receive differential treatment compared to opposite-sex couples in public accommodations (Jones 1996; Walters and Curran 1996). While these studies are suggestive, they tend to be severely under-powered, show small effects that are inconsistent across outcome measures, or involve treatment conditions that strain credulity.\(^2\)

\(^2\)In Hebl et al. (2002), for example, volunteers were sent to apply for jobs wearing hats that read “gay and proud.” In other studies, fictitious resumes include work histories that may signal both sexual orientation and variation in...
More importantly, this existing research focuses on private discrimination and was conducted prior to important changes in American politics. Most audit studies of sexual orientation discrimination were published in or before 2003, around the time that same-sex marriage became the central organizing issue of the LGBT rights movement in the United States. Since then, studies have documented changes in public attitudes about LGBT people and their rights (Pew Research Center 2013, 2017). Moreover, there have been significant changes surrounding the legality of same-sex marriage in the United States. Prior to 2004, most states did not have bans on same-sex marriage written into law, although same-sex couples did not have access to the institution of marriage. After Massachusetts became the first state to legalize same-sex marriage, a social conservative counter-movement led to an increase in the number of states with legal bans from 3 to 30 between 2004 and 2012. Over the same period, the number of states with legal same-sex marriage increased from 1 to 8.

In state-wide referendums in 2012, Washington, Maryland, and Maine voted to legalize same-sex marriage and Minnesota rejected a ban on same-sex marriage. One year later, the Supreme Court ruled part of the Defense of Marriage Act unconstitutional in United States v. Windsor (2013), granting legally married same-sex couples equal status under federal law. Between 2013 and 2015, the number of states with bans on same-sex marriage decreased to 14. These final 14 states were eventually forced to legalize same-sex marriage when the Court ruled all bans on same-sex marriage unconstitutional in Obergefell v. Hodges (2015).3

Nonetheless, de jure extension of marriage rights in the United States did not end political controversy surrounding the rights of same-sex couples. Since the Obergefell decision, opponents re-framed the debate about LGBT rights as violations of the freedom of religion. The most famous example is Kim Davis, a Democratic county clerk in Kentucky, who refused to provide a marriage experience.

3 Although states are the actors who define and set marriage laws, the Defense of Marriage Act (DOMA) was enacted at the federal level in 1998. The law passed with veto-proof majorities and was signed by President Bill Clinton after the Hawaii State Supreme Court ruled that bans on same-sex marriage constituted gender discrimination in 1997. Section 2 of the law defined the federal governments position on the legality of same-sex marriage by only affording federal marriage rights and benefits to heterosexual couples. Section 3 of the law allowed states to not recognize legal same-sex marriages that were performed in other states (or countries). Section 2 was ruled unconstitutional in Windsor and section 3 was ruled unconstitutional in Obergefell.
license to a same-sex couple in 2015 on the such grounds (Blinder and Perez-Pena 2015). The state of Kentucky then amended its law to exempt clerks from signing marriage licenses (SB021620.1000 - 1695 - 8234). Similarly, Roy Moore of the Alabama State Supreme Court directed counties to defy the United States Supreme Court ruling (Robertson 2016). As of 2018, there are 8 counties in Alabama that no longer grant marriage licenses, to avoid issuing licenses to same-sex couples (Dunigan 2018). In Texas and New York, same-sex couples have reported being denied marriage licenses by county clerks as recently as July 2018, more than three years after the legalization of same-sex marriage nationwide (Pager 2018; Sanchez 2018). In West Virginia, a couple reported a marriage license granting official had called them an “abomination” (Wooston and Somashekhar 2017). This suggests public officials may be systematically less responsive to same-sex couples. Though this is our primary expectation, we also consider two mechanisms for heterogenous effects.

**Historical Legacy**

First, the weight of evidence in the social sciences suggests that past institutions may have persistent impacts on present service provision. A large and active literature examines the long-term impacts of legal history on economic development (for a review, see Nunn 2011). In American politics, the lasting effects of (defunct) political institutions are felt across a variety of contexts. For example, Acharya, Blackwell, and Sen (2016) argue slavery has lasting impacts on political attitudes, specifically racial animus and conservatism. Another recent example is Trounstine (2018), who demonstrates prior housing institutions impact economic inequality and reinforce contemporary segregation.

There are numerous reasons that suggest laws that discriminate against same-sex couples will have similar effects. Anecdotally, with the exception of New York, all reported cases of marriage license discrimination against LGBT couples occurred in states with overturned same-sex marriage bans or anti-sodomy laws. Moreover, marriage license-granting shares institutional features that are thought to aid in historical transmission. Housing policies, for example, also saw sweeping nondiscriminatory changes to policy ordered via judicial directive—but were implemented by a highly decentralized system of private actors and local governments. Finally, the political context
that created them—including the attitudes of the population from which officials are drawn—
cannot be expected to dissipate overnight.

**Selection Rules and Responsiveness**

Existing audit studies of public officials typically examine contexts with little or no institutional variation. In the case of marriage, however, licenses are granted locally by county clerks, municipal clerks, probate judges, clerks of wills, county registrars, luxury resorts, and centralized state offices. In total, there are more than 6,000 marriage license-issuing authorities in the United States. For our purposes, one key factor to consider is that some marriage license-issuing authorities are appointed, while others are elected. Accountability mechanisms like these are thought to shape the behavior and responsiveness of public officials (Kimball and Kropf 2006). Research has found conditional evidence that local election officials are more responsive to co-partisans in some contexts, although means of selection did not matter to responsiveness (Porter and Rogowski 2018). Other work finds that selection institution shapes the behavior of judges (Canes-Wrone, Clark, and Kelly 2014). When judges are elected, they are responsive to voter ideology (Lim 2013) and alter their sentencing behavior in response to electoral incentives (Huber and Gordon 2004). When their reappointment is conditional on approval from politicians, judges alter their behavior to align with the preferences of the legislators (Gray 2017) and governors (Gray N.d.). Similarly, studies of regulatory commissions have found that selection method influences the behavior of bureaucrats (Besley and Coate 2003; Fields, Klein, and Sfiridis 1997).

This suggests marriage license-issuing authorities may vary in responsiveness depending on their method of selection. Specifically, we expect that elected bureaucrats will be less responsive to same-sex couples in conservative jurisdictions, as these authorities align their behavior with the preferences of their constituents. In contrast, we expect that appointed bureaucrats will respond to same-sex and opposite-sex couples at similar rates, since they do not face electoral incentives to respond to constituency public opinion.
Summary of Expectations

In summary, the sudden change in policy, discretion of officials, anecdotes of discrimination, and long-term public attitude change suggest that same-sex marriage is an ideal case for studying the “street-level” implications of major judicial policy changes. We list our expectations below.

**Discrimination Hypothesis.** Licensers will be less responsive to same-sex Requesters.

**Legacy Hypothesis.** Licensers in states with a recent legal history of limiting LGBT rights will be less responsive to same-sex Requesters, relative to officials in other states.

**Selection Hypothesis.** Licensers who are elected will be less responsive to same-sex Requesters in areas of higher local-level conservatism, compared with other Requesters, relative to elected officials in liberal areas and appointed officials.

The legacy and selection hypotheses imply two and three-way interaction effects, respectively. We discuss power more extensively when contextualizing the results. However, it should be noted that our study will only speak to whether there is evidence of substantively large effects for these more demanding hypotheses.

Research Design

We conducted a randomized experiment with a straightforward, $2 \times 2$ design. The marriage license granting official received an inquiry from a Requester $= \{\text{Female, Male}\}$, inquiring on behalf of their future Spouse $= \{\text{Female, Male}\}$. Table 1 reports this setup. Accordingly, our discrimination hypothesis will be evaluated by comparing responsiveness to $G_1$ and $G_4$, with that of $G_2$ and $G_3$. Other expectations will be tested with difference-in-difference estimates among sample sub-groups. We have no ex-ante expectations about differences across gay and lesbian couples, but our design is capable of measuring them by comparing $G_1$ and $G_4$. 
Since we expect heterogeneous treatment effects, we employ multivariate, continuous blocking to improve balance in the subgroups implicated by our hypotheses (Moore 2012, Moore and Schnakenberg 2016). Specifically, we block based on four covariates: service level (county or city, see Figure 1), institution type (elected or appointed), contact means (email or online form), local ideology, and population. Local ideology is measured with the average Democratic share of the two-party presidential vote from 2008 and 2012. Population estimates were obtained from the Census Bureau. Subnational variation in the level of marriage licensing authorities and the sparsity of population in some counties necessitates blocking on population. Moreover, local population is correlated with ideology because of the influence of urban city centers. Thus, we were concerned about uncovering a spurious relationship between ideology and responsiveness. Urban authorities might be more responsive, for example, because they are more professionalized and accustomed to handling such inquiries.

Sample. Our subject population is 5,123 jurisdictions who issue marriage licenses in the 48 contiguous states. In Hawaii, marriage licenses are granted by numerous private individuals and businesses (hotels and resorts, in particular) authorized by the state. We have not acquired the full registry of these grantors, who number in the thousands. Moreover, since they are not public officials, they are outside the scope of our study. In Alaska, licenses are granted by cities, counties, and the state, and the application process is centralized. The District of Columbia, likewise, operates a single marriage bureau. We exclude both of these outliers. We must also exclude officials who do not have electronic contact information posted online, which leaves 4,414 cases. Contact information either came in the form of an email or an online contact form. Missingness of electronic contact information was correlated with ideology. Notably, we believe we over-estimate
the true number of jurisdictions, as our list is obtained by assuming every city in applicable states (even those with fewer than 50 residents and no webpage) issues marriage licenses. In short, many of the missing cases may be outside the scope conditions of our study. We defer discussion of the influence of these excluded jurisdictions on our analysis to the Discussion and Supplementary Information (SI).

Though the title of the officials who issue licenses varies by state (clerk, recorder, registrar, judge, etc.), the most basic distinction between them is the level of government that issues the license. Figure 1 plots this regional variation—as states in the northeast allow cities to issue licenses.

![Image of the United States map highlighting marriage granting institutions in the northeast.](image)

**Figure 1** – Marriage Granting Institutions in the U.S.

**Instrument.** The experimental instrument is an email inquiry from a fictional constituent (Figure 2). To increase the likelihood that the official received the treatment, it was short and the treatment appeared on the first line. We also avoided declarative statements about identity and randomly vary salutation and email structure to reduce the likelihood of subject discovery (Butler and Crabtree 2017). This addresses a potential concern in prior research showing discrimination against same-sex couples. First names were chosen to provide a clear signal of gender. These were Brandon, Dylan, Jacob, Elizabeth, Jessica, and Megan. In addition, we chose surnames that have consistent interpretation as white across regions, to prevent variation in attitudes about race and ethnicity from influencing our findings (Crabtree and Chykina 2018). These were Anderson, Niels-
son, and Walsh. This also means that the results should be regarded as limited in scope to whites. While we believe there may be important interactions between race and sexual orientation, we leave this for future research for two reasons. First, we have no theoretical priors about heterogeneity extension of LGBT marriage rights by race and ethnicity. Second, power considerations prevented us from fully crossing sexual orientation with race and gender in the design.

Figure 2 – Email Instrument Example

Hello,

My name is [random: {female full name, male full name}]. My future [random: {husband, wife}], [random: {female full name, male full name}], and I need to get a marriage license. When can we apply for one? How long is it valid? How much does it cost?

We plan to tie the knot on [date].

Thank you,
[rand: {female first name, male first name}]

One advantage of the policy chosen for study is that the rules determining cost, expiration, valid applicants, and other features vary by state, and in some cases, locality. Thus, inquiries like these are common, and responses about these details provide a measure of quality. This is important, as the quality of the information provided is arguably as important as receiving a response. In addition, many webpages did not contain this information. In fact, some webpages indicate that marriage licenses are granted via appointment, and thus, require correspondence to schedule. Other officials indicated (via email) that the information provided on their webpage was inaccurate or outdated. In short, the realism of the instrument presents an additional advantage that avoids the pitfalls of asking for information which is always duplicated online.

We sent email messages to all validated addresses and contact forms in late Summer 2018. We selected this time because it was at least a month in advance of the fictitious wedding in September, which is among the most popular months for a wedding in the United States.\(^4\) The emails were sent from six accounts over two days to prevent being marked as spam. The contact forms were manually entered by researchers those same days. The data collection process had a terminal date

\(^4\)Most recent Centers for Disease Control data indicate the most popular months are July, August, and September. URL: https://www.cdc.gov/nchs/data/nvsr/nvsr58/nvsr58_25.htm
on a Saturday in late September 2018, which was the latest wedding date provided in the outgoing emails. Any responses after this date were coded as non-responsive.

No response suggested discovery of the experiment. In fact, in some cases, the lack of a follow up response from the fictitious emailer by the week of the wedding date was alarming enough to provoke additional follow-up emails by officials. There were 4 jurisdictions that had to be dropped for SUTVA violations because different treatments were sent to the same email address by mistake. Yet, even in these cases, officials provided unique responses that suggested they handled the inquiries as though they came from constituents.

**Externalized Costs and Ethical Considerations.** In addition to standard ethical considerations raised in past work, we provide estimates of the externalized cost of the study. It is important to note that the experiment was reviewed and approved by two internal review boards. Still, past research discusses the ethics of deception and lack of informed consent particular to audit experiments of public officials (e.g. Butler and Broockman 2011; Einstein and Glick 2017). Consistent with this research, and because of the routine and informational nature of the inquiry, we argue that our study minimizes the potential harm and burden placed on participants. However, we also argue that providing a more precise estimate of the total burden placed on participants is appropriate. The prominence of the research we cite and the reduced cost of fielding has led to the proliferation of audit experiments of public officials in the social sciences. Potential long term consequences are not difficult to imagine: increased skepticism of constituent communication by public officials, or aggregate costs that exceed the benefit of new knowledge claims.

For the purposes of this study, we define externalized costs as the total public expense siphoned by the experiment—meaning the costs are external to the researcher. The most obvious way to operationalize these costs is to estimate the time our study took away from other work. In this case, we received 3,285 responses. The responses, including all salutations and email signatures contained 460,433 words, or about 140 words per reply. Estimates of average typing speed in the U.S. suggest each of these replies took about 3 or 3.4 minutes on average. If we use the 2017 median hourly wage for license clerks according to the Census, this means a conservative (i.e. ceiling) estimate for the total external cost of the study was $3,355.92, or about 187 hours of work from public officials, nationally. This is likely conservative, since some replies included text
which was part of a template response. With the same simple arithmetic, the cost of reading would be $266.89—assuming all sent emails that did not bounce were read. Of course, these are not invoices. By providing them, we argue that researchers conducting similar experiments should be precise about externalized costs, and illustrate that minor interventions may have a non-trivial public “burden” in the aggregate.

In this case, we argue the proposed study is worth these externalized costs for several reasons. First, the dearth of research on discrimination against LGBT and the improvements on past designs imply that the simple act of measuring differences is important. Second, the unique historical moment provides an opportunity to learn about additional processes, namely, bureaucratic compliance with judicial directives. In our view, the social value of learning about either phenomena is worth the costs described above. We should also note that since our selection and legacy hypotheses imply covariate interactions, we could not limit the sample of jurisdictions selected for the audit. The reduction in power precluded the possibility of attempting to minimize the number of officials contacted to reduce the externalized cost.

**Response Coding.** We received 3,285 responses from 2,971 unique jurisdictions. We are interested in variation in the responsiveness of public officials to the fictional inquiry. We measure this several ways. We included indicators for whether the official replied and whether the reply included any variation of the word “congratulations.” Neither measure is subject to researcher discretion. The former is a blunt measure of response quality, whereas we argue the latter is a rough proxy for the officials’ attitudes related to the couple in question. Other audit studies use salutations to measure differences across treatments (e.g. Einstein and Glick 2017). In this case, responding with “congratulations” is consistent with American norms in conversations about marriage. Deviation from that norm, then, potentially provides a measure of officials’ attitudes that may be independent from providing factual responses to questions.

Not all responses were of equal quality. Some provide complete answers to every question, whereas others ask the inquirer to find the information on a webpage (without providing the link to that webpage). Some provide information over and above what was specifically asked about, others provide no information at all. To measure this quality, we coded responses for whether they contained responses to the three questions posed in Figure 2: how much it cost, how long it was
valid, and when it could be obtained. Again, there is wide variation by state, county, and city in each of these questions. The lowest cost was $4, the highest was $200. Licenses are valid anywhere from two weeks to, in the words of one official, “eternity.” Some officials require appointments, keep different office hours, or reside in states with variable waiting periods. Some webpages did not contain the necessary information to gauge accuracy. Some respondents indicated that their webpages were out of date. Therefore, we simply code whether each question was answered (not whether it was “accurate”). To check the robustness of congratulatory language, we also had coders determine whether the reply included any felicitations. This included both congratulatory language, and any other well wishes specific to the marriage or wedding.

We used single-blind coding to assess whether the response met the above criteria. Specifically, we removed names of fictitious emailers and all partner synonyms so that coders did not know the treatment condition. Responses were coded by two coders; when coders disagreed, a third resolved the discrepancy. Under this coding scheme, emails that were coded as poor in quality tended to only ask follow-up information, forward the emailer to a webpage, omit answers, and exclude any indication of nuptial excitement. Additional details about this process, along with inter-rater reliability information can be found in SI-B.

Findings

Overall, we find limited evidence of systematic differences across couples, and no evidence of conditional treatment effects by selection or state legal history. Table 2 reports summary statistics by treatment group. The overall response rate of 71% is comparable to other surveys of local-level officials in the U.S. (White, Nathan, and Faller 2015; Giulietti, Tonin, and Vlassopoulos 2015; Dynes, Hassell, and Miles 2018) and Germany (Grohs, Adam, and Knill 2015), and higher than those targeting elected officials (Butler and Broockman 2011) and bureaucrats administering federal programs (Einstein and Glick 2017). To avoid post-treatment bias, all response quality measures (including congratulations rates) are obtained by including both non-response and non-quality as a zero (Coppock 2018).

Among measures of quality, information about cost was provided most readily by officials.

---

5 For a more systematic meta-analysis, see Costa (2017).
Congratulatory language was the least common—but still present in about 1 out of every 6 responses. As is typical in email audit studies, errors and bounces were unevenly distributed across treatment groups, which led to imbalance across one or more covariates. In this case, there were significant differences in population across treatment groups (see Table A1). Given this, we adopt a parametric approach to estimating treatment effects.

Table 2 – Mean Response by Treatments in All States

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Response</th>
<th>Congrats</th>
<th>Cost</th>
<th>Valid</th>
<th>When</th>
<th>Well Wishes</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male Emailer, Male Spouse</td>
<td>0.684</td>
<td>0.124</td>
<td>0.478</td>
<td>0.465</td>
<td>0.328</td>
<td>0.126</td>
<td>1093</td>
</tr>
<tr>
<td>Female Emailer, Female Spouse</td>
<td>0.750</td>
<td>0.129</td>
<td>0.537</td>
<td>0.509</td>
<td>0.388</td>
<td>0.129</td>
<td>1125</td>
</tr>
<tr>
<td>Male Emailer, Female Spouse</td>
<td>0.690</td>
<td>0.137</td>
<td>0.478</td>
<td>0.464</td>
<td>0.353</td>
<td>0.132</td>
<td>1111</td>
</tr>
<tr>
<td>Female Emailer, Male Spouse</td>
<td>0.723</td>
<td>0.139</td>
<td>0.529</td>
<td>0.499</td>
<td>0.379</td>
<td>0.134</td>
<td>1095</td>
</tr>
</tbody>
</table>

Figure 3 plots predicted probabilities for response and congratulatory language by treatment group with 90% and 95% confidence intervals. Models include treatment group, state legal history, service level, population, selection mechanism, Democratic presidential voteshare in jurisdiction, and contact method. We adopt the observed case approach recommended by Hanmer and Ozan Kalkan (2013) to estimate each quantity. Specifically, we estimate first differences after simulating a change in each treatment condition from 0 to 1 for all observations in the data and taking the difference between same-sex and different-sex conditions, while fixing each observation to the observed covariate values. In general, our substantive findings are not sensitive to the covariates included in these models.
Consistent with mean response rates, there are only minor differences across sexual orientation of couples. Officials appeared most likely to respond to female emailers. In general, estimated response rates suggest that number and prominence of female names was more predictive of responsiveness than the sexual orientation of the couple. The difference between same-sex and opposite-sex couples is small in magnitude and not distinguishable from zero (see Figure 7). Moreover, because of the high response rate in the lesbian couple treatment, it runs counter to expectations. Congratulations rates are signed consistent with expectations, but differences for this outcome are again, small and not distinguishable from zero.

As Figures 4 and 5 suggest, these findings are also consistent across various measures of response quality. Answers to questions about license cost, expiration, and application time were all near zero, though inconsistently signed. The lack of differences in congratulations rates is robust to our hand-coded measure of officials’ felicitations. Notably, each quality measure preserves the rank order of responsiveness among treatment groups associated mostly with gender, not sexual orientation. The substantively small, and inconsistent direction of the differences in Figure 3 – Predicted Probability of Response and Congratulations by Treatment Group. Plots point estimates and 90/95 percent CIs simulated from observed data and logistic regression, with controls for state legal history, institution level, population, selection method, democratic presidential voteshare, and contact method.
5 suggests that the results are not an artifact of the particular responsiveness measure. Though we prefer presenting more substantively meaningful outcomes, we should also note there are no differences among aggregated indexes of responsiveness constructed from these indicators.

Figure 4 – Predicted Probability of Quality Response by Treatment Group. Plots point estimates and 90/95 CIs simulated from observed data and logistic regression, with controls for state legal history, institution level, population, selection method, democratic presidential voteshare, and contact method.
Selection Mechanism and Legal History

Though we are less confident in ruling out small to moderate effects, it is important to note that the results were not conditional on either selection mechanism or state legal history. In fact, most of the overall estimates are similar. We estimated differences among treatments for the 26% of all appointed officials (conditional on local ideology) and those who were not. We also estimated first-differences among states with prior legal history unfavorable and favorable to same-sex couples. Specifically, we group officials in states that had unfavorable laws prior to the Supreme Court decisions in *Lawrence v. Texas* (2003) and *Obergefell* (2015) that overturned them. As Figure 6
suggests, these states overlap but are not synonymous. In general, the CATEs we present are not sensitive to the particular measure of state legal history used.

Figure 6 – State Laws Affecting Same-Sex Couples

Figure 7 plots differences across states. Not surprisingly, no difference-in-differences is significant from zero. This consistency is also robust to another potential proxy for attitudes, jurisdiction liberalism measured with Democratic presidential vote share. While we plot results for CATEs for state legal history, the same consistency extends to selection mechanism. According to our

6These legal histories are marginally correlated at 0.26 ($p < 0.1$). It is also worth noting that neither legal history appears correlated with selection method, which suggests our design can plausibly separate the two.
data, the differences-in-differences among appointed and elected officials by jurisdiction conservativism is small and not distinguishable (by conventional thresholds) from zero (see Figure C3 of the SI). Finally, though we report differences for outcomes coded automatically, the results hold for measures of the quality of responses. Finally, the results are also substantively similar for simple difference-of-means tests. Put simply, we find no evidence of substantial differences across sexual orientation in these marriage license inquiries—regardless of jurisdictional subset, local institutions, or measure of responsiveness.
Figure 7 – No Evidence of Responsiveness Differential from Simulated First-Differences. Plots marginal change in predicted probability of response or congratulations, where negative values indicate a lower rate for same-sex couples. “All States” corresponds to experiment ATE, all other estimates are CATE based on state subsets.

Discussion

The experiment did not uncover differences across couple-type that were consistent with our theoretical expectations related to LGBT politics and executive accountability. Our preferred sub-
stantive interpretation of this finding is that anecdotes of LGBT couples facing discrimination are extreme cases, and that judicial change was largely successful at achieving bureaucratic compliance. In general, our experiment suggests that differences in (email) correspondence are either small, undetectable by our measures, or just as likely due to chance. We discuss these conclusions in light of several potential concerns about the study design.

**Power.** Our sample size is comparable to or exceeds other recent audit studies that find evidence of bias and heterogenous treatment effects (e.g. Butler and Broockman 2011; White, Nathan, and Faller 2015; Einstein and Glick 2017; Jilke, Dooren, and Rys 2018; Carnes and Holbein 2018). However, since our results suggest no bias, it is important to contextualize the magnitude of these effects and the power of the design. We argue that the interpretation that there is no evidence of moderate-to-large bias is warranted for several reasons.

First, the conventional power threshold of 80% suggests that our study should uncover effect sizes of 4 percentage point differences (or more) in responsiveness (see Figure A1 in the SI). Importantly, other email audit studies have found more substantial effects for bias against other marginalized communities. For Latino names, for example, White, Nathan, and Faller (2015) find a reduction in responsiveness of 5-7 percentage points, and Einstein and Glick (2017) find differences in email tone of 20 percentage points. Thus, one way of contextualizing our findings is that they suggest bias against LGBT people is likely less than what is typically found for other marginalized groups. Second, and relatedly, we are more confident in these conclusions because our survey instrument generated multiple outcomes that point to the same basic finding. Finally, we find substantial differences by gender, which seem to countervail the hypothesis that sexual orientation drives bias. Couples with two female names consistently exhibited the highest levels of responsiveness.

**Missing Officials.** There is no national registry or list of marriage license granting officials. Our best estimate is 5,123. We arrived at this list by using Census lists of counties in states where officials operated at that jurisdiction, and municipalities in states in which licenses are granted by cities. However, some city officials in small municipalities indicated they had functionally delegated this role to larger cities nearby. In one case, an official indicated they had not received a
marriage license application “in ages.” Outdated city population estimates suggest some municipalities legally permitted to issue licenses do not because there is no official employed to do the job. Other jurisdictions may be counted as cities by the census but are now unincorporated territory. In short, the population of jurisdictions legally authorized to issue licenses is likely different than those who actually issue licenses.

Moreover, some jurisdiction’s webpages indicate they issue licenses, but provide only a physical address or phone contact information. Other jurisdictions had no web presence. These officials were excluded and tended to be in smaller, more rural and politically conservative localities. Our analysis, therefore, cannot speak to the behavior of these officials, and it is possible that their inclusion could change the results. We discuss this further in Appendix A of the SI. Notably, there was no relationship between any of these covariates and our measures of differences in the larger sample. Thus, we do not expect that including them would change our conclusions.

**Treatment.** Our treatment was designed to closely resemble an typical email inquiry these officials would receive. It requires the official to make an inference about sexual orientation based on names and a gendered partner synonym. One potential explanation for our finding is that the magnitude of the treatment was insufficient to be received or to prime the expected response. But the subtlety of this treatment mirrors of that numerous email audit studies that did uncover evidence of bias. This, of course, does not rule out the possibility of systematic differences in face-to-face or phone interactions, which would require far more resources to study properly.

**Conclusion**

Our study provided an experimental test of bureaucratic responsiveness to same-sex and opposite-sex couples requesting information about obtaining a marriage license. In contrast to other audit studies investigating discrimination against marginalized groups, we find little to no evidence of this toward LGBT couples seeking a marriage license. Our findings have important implications for research on LGBT politics and executive accountability, more broadly.

---

7One example is Centerville, ME, whose fewer than 50 residents voted to dissolve the town government after the 2000 census.
First, our study examines the implications of a policy recently settled by the Supreme Court. Historical moments in which bureaucrats must implement far-reaching policy change are necessarily rare, which sets our analysis apart from other audit studies. Our null results coupled with the timing of our study, just 3 years after 14 states were forced to legalize same-sex marriage, suggests that executives, who have discretion to implement law, have largely complied with judicial directives related to same-sex marriage. This contribution is best situated in broader historical context surrounding the implementation of civil rights for minority groups. In other cases of civil rights, bureaucratic officials slowed the implementation of civil rights for African American and under-represented minorities. When viewed alongside these other civil rights examples, our findings suggest that discriminatory biases toward marginalized groups vary across group and policy context.

We are not suggesting that LGBT people no longer face discrimination in the United States. LGBT people continue to report experiences of discrimination at high rates (Pew Research Center 2013). In 2017, the Human Rights Campaign documented 129 anti-LGBT bills introduced in states (Warbelow and Diaz 2017). President Trump has implemented a ban on transgender people’s service in the military (Marimow 2019). Moreover, our findings do not speak to more subtle bias not typically detectable in a study of this size. At a minimum, however, our study aligns with other evidence that shows improving social and political acceptance of LGBT people. Finally, our study is consistent with emerging evidence that suggests bureaucratic agents concerned chiefly with service provision on a day-to-day basis exhibit less evidence of bias than politicians (Einstein and Glick 2017; Porter and Rogowski 2018) and private actors (Jilke, Dooren, and Rys 2018).

Our study also suggests avenues for future research. For scholars of LGBT politics, the difference in bureaucratic responsiveness to gay and lesbian couples is particularly noteworthy. This finding is similar to research that has found gay men receive lower evaluations compared to lesbians in other contexts and more work needs to examine the foundations of these gender and sexuality based biases. Importantly, this is largely driven by the finding that bureaucrats were more responsive to female requesters in general. Since we did not have initial expectations that our results would be driven by requester gender, we do not engage in post-hoc analyses to better understand this finding. One possibility is that women may be more likely to serve in public office as license issuing authorities and we could be observing higher responsiveness to other women. This
would be consistent with findings of responsiveness to in-group constituents (Butler and Broockman 2011). Future research can address this topic, contributing to a growing literature about the representation of women in local government (Holman 2017; Einstein, Palmer, and Glick 2018). Finally, our findings are limited to understanding discrimination against white LGBT people and future research should explicitly examine discrimination and biases at the intersection of race, sexuality and gender. Such an agenda might help scholars better understand the complexities and contingencies of discrimination in the United States.

References


Supporting Information (Online)
Bureaucratic Responsiveness to LGBT Americans

A  Study Diagnostics  A2
A.1  Power  A2
A.2  Missing Jurisdictions  A2

B  Response Coding Procedures  A4

C  Results for Institutions Hypothesis  A5
Study Diagnostics

Power

We simulated design power for each hypothesis with the observed $N$ and distributions of each interactive variable. These simulations differ from the analyses presented in the paper, as they assume simple random assignment and exclude additional covariates. Figure A1 plots mean proportion of 1,000 simulations which return significant (i.e. $p < 0.05$) differences in the expected direction by effect size. The mean for the dichotomous outcome is set to 0.71, the value observed in the data. By conventional power thresholds, this design was well-powered to detect substantively moderate effects (±4%) for the overall discrimination hypothesis, large effects (±8%) for the legacy hypothesis, and under-powered for the final subgroup effects. Accordingly, our experiment does not provide evidence in support or in-favor of substantively smaller effects in either the overall sample or in the theoretically-informed subgroups.

![Figure A1 – Power Analysis](image_url)

Missing Jurisdictions

What jurisdictions were less likely to be included in our sample? Figure A2 plots missing counties and the proportion of missing cities in municipal-granting states. In total, there were roughly 704 jurisdictions without online contact information. The exact number is unknown for reasons we elaborate on in the Discussion. While there appears to be no relationship between population and inclusion, excluded jurisdictions tended to be more conservative. In models controlling for population, a standard-deviation increase in two-party Democratic presidential vote share from 2008 and 2012 is associated with a 3.5% increase in the probability of inclusion. (The unconditional probability of inclusion was 86%.) Note also, conservatism is correlated with census designations for “rural” geographic areas.

A2
During the design stage, we expected lower responsiveness, weaker web presence, and higher conservatism would co-occur. Since the study found no CATEs among various levels of jurisdiction conservatism, we have no reason to expect that the inclusion of the missing jurisdictions would alter the results. Moreover, if these jurisdictions were included, they would account for a negligible proportion of the marriage licenses granted in the U.S., so the substantive interpretation of our study would likely be unchanged. Nonetheless, future studies should note this limitation, as it may be addressed by conducting an audit via telephone.

Figure A2 – Missing Jurisdictions
Table A1 – Imbalance Across Treatment Conditions

<table>
<thead>
<tr>
<th></th>
<th>Male, Male</th>
<th>Female, Female</th>
<th>Male, Female</th>
<th>Female, Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>59519.59</td>
<td>53767.62</td>
<td>66697.34</td>
<td>60029.81</td>
</tr>
<tr>
<td>Democratic Voteshare</td>
<td>0.45</td>
<td>0.45</td>
<td>0.46</td>
<td>0.45</td>
</tr>
<tr>
<td>Appointed</td>
<td>286.00</td>
<td>286.00</td>
<td>286.00</td>
<td>286.00</td>
</tr>
<tr>
<td>Contact Forms</td>
<td>58.00</td>
<td>100.00</td>
<td>102.00</td>
<td>81.00</td>
</tr>
<tr>
<td>Errors</td>
<td>81.00</td>
<td>117.00</td>
<td>122.00</td>
<td>97.00</td>
</tr>
<tr>
<td>Bounces</td>
<td>118.00</td>
<td>150.00</td>
<td>158.00</td>
<td>143.00</td>
</tr>
<tr>
<td>N</td>
<td>1093.00</td>
<td>1125.00</td>
<td>1111.00</td>
<td>1095.00</td>
</tr>
</tbody>
</table>

Response Coding Procedures

Cost, waiting period, and time valid can vary by unit. Many webpages did not contain the necessary information to gauge accuracy. In addition, some respondents indicated that their webpages were out of date. Therefore, we simply code whether each question was answered. We used single-blind coding assess whether the response addressed the questions posed by the inquiry. Specifically, we followed the procedure below:

1. Removed treatment condition identifying information, and all names and partner synonyms from the text of replies,

2. Randomized the order of response coding by rater.

3. Coded each of the 4736 observations according to the following rules:

   **Cost.** “Did the official answer the question about the cost of a marriage license?” An exact amount had to be provided. If the official indicated that the cost is dependent on residency or premarital counseling, they must provide costs for with/without each status to be counted as responsive.

   **Valid.** “Did the official answer the question about how long the marriage license was valid?” Official must indicate if the license expires, and when. Replying that the emailer can “apply anytime” is not sufficient.

   **When.** “Did the official answer the question about when you could apply for a marriage license?” If they said “you can apply anytime now (e.g. there was no statutorily mandated waiting period) it must include office hours to be counted as responsive. Thus, either notification of office hours or information about a waiting period iss considered responsive.

   **Felicitations.** “Did the official say ‘congratulations’ or wish ‘good luck’ at the wedding/marriage, etc.?“ Email signatures and salutations like “best” or “have a nice day” are not counted as a felicitation.

4. Disagreements between raters were resolved by a third, double-blind coder—who was unaware of the study purpose.
Table B2 – Interrater Reliability for Response Quality

<table>
<thead>
<tr>
<th></th>
<th>Cost?</th>
<th>Valid?</th>
<th>When?</th>
<th>Felicitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Disagree</td>
<td>0.01</td>
<td>0.03</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>No. Disagree</td>
<td>47.00</td>
<td>143.00</td>
<td>226.00</td>
<td>112.00</td>
</tr>
<tr>
<td>Cohen’s Kappa</td>
<td>0.98</td>
<td>0.94</td>
<td>0.89</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Table B3 – Interrater Reliability for Response Quality (Non-Response Removed)

<table>
<thead>
<tr>
<th></th>
<th>Cost?</th>
<th>Valid?</th>
<th>When?</th>
<th>Felicitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Disagree</td>
<td>0.02</td>
<td>0.06</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>No. Disagree</td>
<td>47.00</td>
<td>143.00</td>
<td>226.00</td>
<td>112.00</td>
</tr>
<tr>
<td>Cohen’s Kappa</td>
<td>0.89</td>
<td>0.76</td>
<td>0.80</td>
<td>0.87</td>
</tr>
</tbody>
</table>

In general, coding was consistent across raters to suggest that a third coder to resolve discrepancies was sufficient. Not surprisingly, cost was mostly clear for human coders, whereas the most disagreement arose for the more generic question of “when” the emailer could apply and whether the reply included some kind of felicitation.

Results for Institutions Hypothesis
Figure C3 – No Evidence of Differences in Response Rates (Simulated) by Institution and Ideology. Plots marginal change in predicted probability of response quality, where negative values indicate a lower rate for same-sex couples. For the purposes of this figure, liberal and conservative is defined as 35 and 65 percent Democratic two-party voteshare in the 2008 and 2012 presidential elections.