An Introduction to Conducting Email Audit Studies

Charles Crabtree*
University of Michigan
ccrabtr@umich.edu

April 24, 2017

Abstract

This chapter offers the first general introduction to conducting email audit studies. It provides an overview of the steps involved from experimental design to empirical analysis. It then offers detailed recommendations about email address collection, email delivery, and email analysis, which are usually the three most challenging points of an audit study. The focus here is on providing a set of primarily technical recommendations to researchers who might want to conduct an email audit study. The chapter concludes by suggesting several ways that email audit studies can be adapted to investigate a broader range of social phenomena.

*I thank Volha Chykina and Christopher J. Fariss for their helpful comments. I particularly thank Holger L. Kern for teaching me about audit studies and providing me some of the code used to conduct email audit studies.

1 Introduction

What is an audit study? As other chapters in this volume note (Gaddis, 2017*b*), an audit study (or correspondence study) is one way of assessing hard-to-observe phenomena, such as discrimination (Heckman, 1998).[1] The general structure of an audit study is very simple. To begin with, researchers create some set of identities. The initial identities share the same characteristics. Scholars then randomize one or more attributes of the identities, such as race or gender. Next they use these identities to accomplish some task, like applying for jobs, renting housing, or contacting legislators. These tasks can done via phone, mail, and email. Finally, scholars compare how individuals — such as prospective employers, landlords, or legislators — respond to the putative identities. Any difference in treatment across the randomized attributes is interpreted as evidence of some latent bias. For example, if landlords respond to inquiries from Blacks less frequently than inquiries from Whites, then scholars would infer that landlords are biased again Blacks. Scholars have used audit studies to observe biases in nearly every facet of common life — in political interactions (Butler, 2014; Broockman, 2013; Butler and Broockman, 2011; Grose, 2014; Costa, N.d.), in housing transactions (Gaddis and Ghoshal, 2015; Turner et al., 2002; Hogan and Berry, 2011; Oh and Yinger, 2015), in economic exchanges (Riach and Rich, 2002), in employment decisions (Neumark, Bank and Van Nort, 1995; Bertrand and Mullainathan, 2004), and in many other spheres (Pager and Shepherd, 2008). Taken together, the results from these studies have considerably improved our collective understanding of discrimination.

The important point for this chapter is that an increasing number of these audit studies are being conducted over email.[2] There are several reasons for this. One reason is that email is an extremely common means of communication; approximately 2.6 billion people

[1] See Gaddis (2017*b*) for a history of audit studies and an overview of the approach.

[2] Some recent examples of this include Gaddis (2014); Gaddis and Ghoshal (2015); Sharman (2010); Radicati and Hoang (2011); Oh and Yinger (2015); Milkman, Akinola and Chugh (2015, 2012); Lahey and Beasley (2009); Hogan and Berry (2011); Giulietti, Tonin and Vlassopoulos (2015); Findley, Nielson and Sharman (2015); Bushman and Bonacci (2004); Butler (2014); Ahmed, Andersson and Hammarstedt (2012, 2013); Baert (2016); Baert, De Visschere, Schoors, Vandenberghe and Omey (2016); Baert, Norga, Thuy and Van Hecke (2016).

sent over 205 billion messages in 2011 (Radicati and Hoang, 2011). Email can be used to accomplish virtually any communication related task — from exchanging documents, to sharing personal news, to organizing collective actions, to conducting business transactions, or even to requesting assistance from public officials. The dominance of email as a mode of communication is indicated by the fact that workers report spending up to 50 percent of their day reading, writing, and managing emails (Stocksdale, 2013). This widespread use of email helps researchers because it provides them with opportunities to engage in many different types of interactions and thus potentially observe discrimination (or other phenomena) across many contexts. Another reason relates to external validity. As an increasing number of interactions occur over email, researchers would limit the generalizability of their findings if they only conducted audit studies through other media.

A third reason why the number of email audit studies is increasing is because they are relatively inexpensive to implement. There are costs to conducting audit studies through other means, such as the mail, that simply do not apply to email studies. For instance, in the case of mail, these costs might include stamps, post office boxes, enumerators in different locations. In contrast, anyone with an Internet connection can send and receive emails for free. This means that researchers with limited resources — such as graduate students and junior faculty — might find email a particularly attractive means of conducting their correspondence studies.

Despite these advantages, email audit studies are perhaps underused. Certainly, many audit studies have been conducted over email since electronic messaging became widely available. This number could be even higher, though, as every published audit study suggests (implicitly or explicitly) a large number of possible extensions and adaptations.

One reason why email audit studies might not be used more is that they are often difficult to implement. This is particularly true for scholars who are inexperienced with conducting audit studies in any medium. The issue here is that there are no general introductions to audit studies. Another reason why email audit studies might be underused is because scholars

might think that they can only examine a narrow range of social phenomena. While the vast majority of email audit studies have focused on unearthing evidence of discrimination, this general form of study can be easily adapted to examine a wider range of social phenomena.

In this chapter, I address both of these issues with the goal of increasing email audit study use.[3] The first section of the chapter attempts to reduce the complexity of email audit studies by providing a comprehensive guide to implementing one. This guide describes the steps involved in conducting an audit study. It also offers detailed recommendations about how researchers should collect, send, and code emails, since these are perhaps the most intimidating steps to inexperienced scholars. The primary focus of this section is on describing computerized, time-saving solutions to common issues. The `R` code used to address these issues is available online at charlescrabtree.com/email_audit and at auditstudies.com.[4]

The second section of the chapter offers several suggestions about how scholars can adapt audit studies to investigate a broader range of social phenomena. It provides examples of non-traditional audit studies and discusses how those designs might be modified to answer other theoretical questions. This deconstruction of prior research might be helpful to scholars who are interested in audit studies but think that they cannot be used in their research.

2 Guide to Implementation

How can a researcher conduct an email audit study? This section addresses that question by providing an overview of the implementation process. Before discussing individual steps in

[3] I acknowledge that there are instances in which researchers cannot or should not implement an audit study over email. Perhaps the biggest reason for this is it might be impossible to collect email addresses for some populations. For instance, it would be very difficult to get email address information for a random sample of Americans. Similarly, one can imagine international contexts, such as many emerging market economies, where it might even be difficult to gather email addresses for public figures, such as government members. In addition to this concern, it is also probably true that some interventions are less plausible over email than through the regular mail or via phone. To the extent that researchers want to maximize the ecological validity of their interventions, they might want to conduct them via alternative means. Yet, despite these limitations, I still think that there are substantial opportunities for conducting additional email audit studies. These opportunities will continue to increase so long as email remains one of the most widely used means of communication.

[4] While I focus on using `R` to address some implementation issues, researchers should be able to accomplish similar tasks in Stata or using other programming languages, such as Python.

detail, we first provide a general outline of the stages involved in a typical email audit study. These eight stages are listed in Figure 2. They include (1) experimental design, (2) sample selection, (3) email address collection, (4) covariate collection, (5) treatment randomization, (6) treatment (i.e. email) delivery, (7) outcome collection, and (8) analysis.

One additional stage not discussed here is getting institutional approval, typically provided by an institutional review board (IRB), for conducting the intended study. Since the requirements of these boards vary considerably across institutions (Driscoll, 2015), it is difficult to provide useful, general recommendations about how to manage their potential concerns. Interested readers should consult Hauck (2008), Fujii (2012), and Yanow and Schwartz-Shea (2008) for overviews of potential IRB-related problems and solutions. More generally, researchers should carefully consider the ethics of their experimental interventions. Desposato (2015) and Riach and Rich (2004) provide great introductions to these issues, and Gaddis (2017*a*) offers several suggestions regarding best practices.

2.1 Experimental Design and Sample Selection

While experimental design and sample selection are extremely important, I do not discuss them here. Many excellent texts deal with issues related to design and sampling (e.g., (Gerber and Green, 2012; Lohr, 2009). I refer interested readers to them.[5]

One important issue related to sampling stands out — power calculations. These calculations are used to determine whether experiments are sufficiently powered to detect treatment effects. In other words, they help scholars determine if they have included a sufficient number of participants. The Evidence in Governance and Politics Group provides a useful guide (`goo.gl/HXOK5Q`) and a couple calculators (`goo.gl/CJ8zox`, `goo.gl/0ucE9G`) that researchers can use to think through their potential statistical power concerns.

[5]For ease of exposition, I assume that researchers are implementing a between-subjects design. The general process described in this chapter can be easily adapted to accommodate a within-subjects design. The only potential difficulty in doing this would be in modifying the email delivery script available in the online appendix. I have addressed this issue by modifying the code to deal with both types of design.

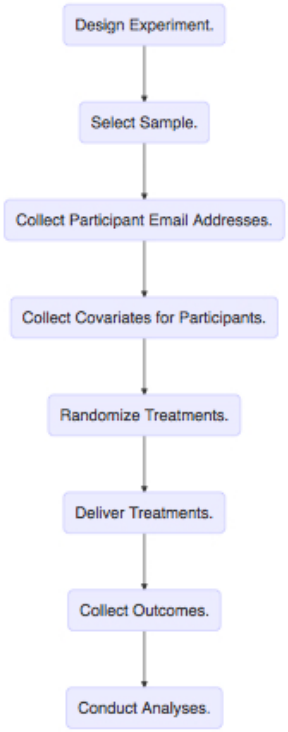

Figure 1: The eight stages of a typical email audit study.

Regardless of what researchers decide regarding experimental design and sample selection, they should consider pre-registering these choices, along with their theoretical expectations and analytic strategy (Olken, 2015; Franco, Malhotra and Simonovits, 2014).[6] There are many possible reasons to write a pre-analysis plan.[7] If scholars pre-register their research designs, they might think more clearly about their theoretical expectations and the extent to which their proposed design might satisfactorily test them. Pre-registration should also lead to fewer questionable research practices, such as analyzing the data in whatever way leads to statistically significant results (i.e. 'p-hacking') or hypothesizing after results are known (i.e. 'HARKing'). This is because it forces researchers to commit to analyzing and discussing the results as discussed in the pre-analysis plan (Olken, 2015). Finally, researchers might want to pre-register their designs because journals in some fields, such as political science and psychology, are increasingly encouraging this practice. Pre-analysis plans can be posted on sites like the AEA RCT Registry, the Evidence in Governance and Politics site, or on personal academic webpages.

2.2 Email Address Collection

Once a researcher has designed an experiment and selected a sample, they need to collect email addresses for each participant in their sample. This is typically one of the most difficult and time-consuming steps in conducting an email audit study. One of the things that make this so difficult is that researchers often want to recruit a large number of participants. This could be because they want to maximize statistical power or because they want to increase the external validity of their findings. Regardless of the reason, gathering contact information and other details for large samples can be intimidating. I briefly discuss here some of the ways that researchers can efficiently collect contact information for their sample.

Thankfully, this task is now perhaps easier then ever before. In many cases, researchers

[6] Lin and Green (2015) provide excellent guidance on some of these decisions.
[7] Coffman and Niederle (2015) discusses some of the limitations of pre-analysis plans.

can find participants' emails online, either individually or together as part of a mailing list. This is particularly true in the case of public figures. Sites like everypolitician.org and sunlightfoundation.com provide data for elected officials. Lists of unelected officials emails are often available from offices in Washington, D.C. or at state capitals.

Even when the information has not already been previously compiled by others, researchers still have many tools at their disposal that can reduce the time they would spend on data collection. One quick way to collect contact information is by scraping it from websites, such as job boards, or state agency employee listings. Building a web scraper used to be something that only a well-trained programmer could manage, but the diffusion of programming tutorials and the ready availability of example code at sites like github.com or stackexchange.com, have made it so that even individuals inexperienced with programming can adapt existing scrapers to their own purposes.

Some sites present problems to basic scrapers, though, such as login screens or paywalls. In these cases, researchers have two options. If they have research funds, they might consider paying a programming freelancer to create a custom scraper for them. Sites like elance.com and guru.com can help researchers find qualified help. Since building a scraper is a rather basic programming task, the job would not cost much. If researchers, however, cannot (or will not) pay for a freelance programmer to build a scraper, then they can explore what-you-see-is-what-you-get solutions, such as the excellent Web Scraper extension for Chrome.

After collecting emails, researchers should drop obviously invalid email addresses. This includes emails that do not contain an '`@`' symbol, emails that contain spaces, and emails that are actually website addresses, among others. One reason to drop bad email addresses before implementing the experiment is to reduce the number of invalid email notifications received post-implementation. Scholars should not worry too much about catching every invalid address, though. Since treatment is randomized, they should be able to drop observations that contain bad contact information without biasing inferences.

2.3 Covariate Collection

Researchers might gather covariates on their participants either prior to or alongside email addresses. There are two general reasons to collect covariates related to their sample. One is to examine treatment effect heterogeneity. This is the "degree to which different treatments have differential causal effects on each unit" (Imai, Ratkovic et al., 2013, 443). The idea here is to use pre-treatment covariates to determine the effect of treatments on different subpopulations. Another reason to collect covariates is to include them in the randomization scheme, such as through block randomization (described below) (Suresh, 2011). In many cases, scholars can use the same techniques to collect covariates as they do to collect email addresses.

2.4 Treatment Randomization

After collecting covariates, researchers should then decide how they intend to randomize treatment. There are many ways that you can do this. One approach would be to just use a random number generator. A more sophisticated approach would be to assign treatments within blocks. This is done by dividing subjects into homogeneous blocks and then assigning treatments within those blocks. The goal here is to increase efficiency by decreasing variability between units. When randomizing this way, I typically use the `R` package `blockTools` (Moore and Schnakenberg, 2012). The choices that researchers face at this step are not unique to email audit studies, though, so I do not discuss them at length here. Gerber and Green (2012) offer a particularly good guide to the pros and cons of various randomization schemes.

2.5 Email Delivery

After scholars randomize treatment assignment, they need to assign those treatments to participants. Since this chapter focuses on email audit studies, I assume that treatments are

being delivered via email. In order to assign treatment then, researchers need to email study participants.

Researchers can send emails manually. This would involve sending each email one-by-one through an email client or web application, such as gmail.com. There are two problems with this approach, though. The first is that it can be time-consuming to send many emails this way. It might also be impractical for researchers who intend to contact very large samples (Butler and Crabtree, 2017). The second is that researchers might make mistakes when sending emails manually. They could, for example, assign the wrong treatment to a participant, or accidentally fail to send emails to some participants. This is a problem because mistakes such as these could lead to invalid inferences.

Researchers can also send emails automatically with the help of a programming script. There are several advantages to sending emails like this. The first is that it can dramatically reduce the time that researchers spend actually sending emails. Instead of addressing emails to individual participants, scholars would only need to execute a loop of code that would iteratively email each participant. The second is that it reduces the possibility of error. If prepared properly, the script should correctly assign treatments and email all participants. A third advantage is that a script can record the exact time that emails are sent. This is useful if scholars have theoretical expectations regarding how treatments influence not only whether individuals respond but how long they take to respond as well. Taken together, these advantages suggest that scholars should send emails through scripts.

While researchers might understand *why* they should do this, it is often less clear about *how* they should do this. I provide a detailed outline of this process below. This is based on a set of best practices developed over more than a dozen email audit studies with various collaborators. The outline is broken down into two sections. The first describes the steps researchers should take prior to sending emails. The second describes the steps involved in sending the emails.

2.6 Pre-Implementation

To begin with, researchers should create an email delivery account for every putative identity used in the experiment. In the past, I used free email accounts from services like gmail.com and yahoo.com. Many free email providers have changed their security policies, though, making them potentially untenable solutions for researchers who want to quickly send their emails through programming scripts. One potential workaround is to modify the script so that it pauses between email sending attempts.[8] Scholars who want to use these services should check their security policies before implementation.

Recently, I have used Google Apps to send email, though other domain hosting services like dreamhost.com would work. While this approach imposes a marginal monthly cost ($5 to $10 a month), it allows scholars to get around the security restrictions now common with free accounts. The main downside of this approach is that it requires emails be sent from a domain name that the researcher registers. In several experiments, I have registered and used domains that include a combination of the first and last name for a putative identity. The potential problem with this, however, is that individuals who send emails from custom domains are presumably different from other individuals in important ways. For example, they probably possess higher tech skills and they might have more disposable income. Another option is to register a domain name for a dummy corp (e.g., `dummy-corp.org`) or email provider (e.g., `thefastestemailever.org`). In order to make the domain name seem more legitimate, I typically put up a basic webpage at that domain. The trick with this approach is that it can be difficult to register domain names that do not bring to mind specific association(s).

Another potential problem with using a custom domain name is that it might raise participant suspicions. This could increase the risk of experiment discovery. Just as worrying, it could also cause participants to behave in ways other than usual. Unfortunately, there is not a clear solution to this problem, and researchers simply have to evaluate the advantages

[8] I provide an annotated example of this in the online appendix for this chapter.

and disadvantages of each email sending approach within the context of their experiment. Regardless of how they decide to send email, they will need to think carefully about the problems their method might pose to the interpretation and external validity (i.e. scope conditions) of their findings.[9]

After researchers have created the email accounts they will use in their experiment, they should create an additional email account. This will be the master account from which researchers can monitor initial responses and collect final outcome data. All email delivery accounts should be set to forward email to this account.

There are three primary reasons to create a master account. The first is that researchers might want to monitor emails as they arrive, so as to make sure that the experiment was successfully implemented. Researchers should avoid monitoring the original replies, though, as it is very easy to accidentally respond to a message. In some cases, a reply might raise participant concerns and lead to unnecessary problems. The second reason is that it is easier to collect outcome data from one account than many. The third is that bad things can happen with email accounts. Researchers can, for example, be locked out of accounts. It is therefore wise to keep multiple copies of the emails across accounts. Since the master email account will only be used to receive emails, I often create a gmail.com account. This is because Google provides an easy interface for exporting emails.

Once researchers have setup the email delivery and master email accounts, they can attend to other details. They need to write the code that links treatment assignments to strings of treatment text, such as the name of the sender. Scholars should also create the strings of text that comprise the non-random email components, such as email valedictions or salutations.[10] After that, scholars will need to write the code that combines the random and non-random strings of text into a complete email. The online appendix for this chapter

[9] Pedula (2017) discusses some of the other issues that potentially limit the generalizability of audit study findings.

[10] In some cases, researchers might want to randomize the valedictions or salutations. This could be a good idea if scholars are concerned about some actor observing similarities across delivered emails (Butler and Crabtree, 2017).

includes `R` code for both steps.

Next scholars should create a script that will deliver their emails. The script should loop through each observation in the dataset. In each iteration, it should extract an observation's email address and treatment details, combine the treatments and other text elements into a complete email, and send the email. After sending the email, the script should save the time that it was sent. This information can be used to confirm that individual emails were sent. It can also be used to create a 'time to reply' outcome measure, as I discuss later. After that, the script should print the observation number for that iteration. This is for diagnosing potential problems later. The online appendix for this chapter includes `R` code for this loop. It is highly annotated and can be easily adapted to fit a variety of needs.

The final step before implementing the experiment is to test the script. I suggest that researchers do this by sending a limited run of emails (20 or 50) to all project collaborators. The idea here is to test all of the email settings saved in the script. An additional benefit of doing this is that everyone working on the project can look carefully through the sent emails. Particular attention should be paid to the email headers and subject lines, which can be easily ignored. If these emails look good, then the experiment is ready to implement.

3 Implementation

Researchers begin implementation by executing the script. In an ideal world, the script will execute successfully, only finishing when all emails are sent. Unfortunately, the script will most likely fail at some point, causing the loop to stop. This can happen because an invalid email address remains in the dataset. Most scripts will be unable to parse invalid email addresses and will register an error when reading them. Since the script prints the observation number at the end of each iteration, researchers can manually inspect the dataset to see if the error was caused by an invalid email. If researchers cannot fix the email address, they then should skip that iteration of the loop.[11]

[11] I have assumed here that all emails can be delivered in a single wave. This might not be possible depending on the email solution used and the size of the participant pool. One potential problem here is that some

The script can also stop because of email server problems. Sometimes servers, even gmail.com servers, are unable to accept email commands. Sometimes servers will only take so many email commands within a short period of time. In either case, the script available in the online appendix will register a server error. The best way to deal with this problem is to wait a few minutes and restart the loop at the current iteration.

While the script is running, researchers should open the master email account and monitor it for responses. Unless the emails are sent at a really odd time, the participant pool is really small, or the requests will take a while to address, responses should pour in shortly after the script has been executed. There are several reasons to check the responses. The biggest reason is to ensure that the experiment was successfully implemented. Evidence for this can come from email replies, which often include the full text of the sent email. Another reason is to ensure that participants appear unaware that they are part of a study.

4 Outcome Collection

Having sent emails, scholars can begin collecting outcomes measures. While audit studies make use of many different outcomes, the primary outcome of interest in many *email* audit studies is a binary indicator that is coded 1 if participants replied and 0 otherwise (e.g., Butler (2014), Bertrand and Mullainathan (2004), and Grose (2014)). There are two ways that scholars can construct this indicator. The first and most common way of collecting this outcome is to read and manually code email responses. The benefit of this approach is that it can be very accurate compared to automatic coding. The problem, however, is that it can be extremely time-consuming to process a large number of emails. Given a sufficiently large sample, it might simply be impractical to do so.

The second way that scholars can collect this outcome is by using a script to automatically code replies. This approach has the benefit of speed, as a script can code thousands of emails

servers might limit the number of emails sent in any given 24-hour period. If researchers need to send emails across multiple waves, they will then need to subset their data into different waves prior to implementation and then execute the script for each wave.

in minutes. The disadvantage of this approach, however, is accuracy. In some cases, emails might not be accurately matched with observations. Most of the time this loss in accuracy is relatively trivial, influencing only a small number of observations.

Before using a script to code emails, scholars first need to download the data from the master email account. The exported data will likely be in `.mbox` format. At this point, scholars could either use the script available in the online appendix or one that they create. The heavily annotated `R` script performs a number of functions. First, it converts the `.mbox` file into N `.eml` files, where N represents the number of email replies. Second, it reads the emails. Third, it extracts the email addresses that are included in each reply. Fourth, it matches those email addresses to observations in the dataset, link email reply and participant information. Fifth, it creates the outcome measure for each observation.

While a binary *email reply* indicator might be a suitable outcome measure for many research questions, scholars might also be interested in other outcomes. For instance, researchers might want to code whether the replies they receive are positive or negative. This would be easy to do manually. Researchers, however, could also do this automatically. The key here would be to identify words and phrases that are unique to positive or negative replies. Once this is done, scholars could adapt the script discussed above to search for these terms within the email texts that have been linked to participants. If the email contains one or more of the words that uniquely identify positive replies, then an observation can be coded as receiving a positive response. An example of how to do this is included in the script.

They might also have theoretical expectations about how treatments influence *when* participants reply. In this case, they might want to record the time participants take to reply. The `R` code included in the online appendix can be easily adapted to extract this information from the email replies. Once researchers know when they received email replies, they can subtract the email sent time recorded in the delivery script from this value.

Scholars might also be interested in the length of replies. Reply length could, for instance,

be used as a measure of email helpfulness. While scholars can count the words in each reply, it is much easier to do this automatically using either the included code or commercially available software, such as Linguistic Inquiry and Word Count (LIWC) (Pennebaker, 2015).

Finally, researchers might be interested in examining the sentiment of the replies. For example, they could be interested in how positive or negative the replies were. Scholars could create this measure manually, by reading and assessing each email. Or they could use one of several software solutions. For example, LIWC can generate measures of positive and negative emotion (Pennebaker, 2015). The difference of these two quantities can be taken as a measure of positive sentiment (Crabtree et al., N.d.). Another way that researchers can code this measure is through natural language processing (Manning et al., 2014).

5 Analysis

Once scholars have collected their outcomes of interest, they can analyze the results. There are good guides for analyzing experimental results, such as Gerber and Green (2012). For any additional data analysis needs, I recommend Gelman and Hill (2006).

6 Extending Audit Studies

Having explained how scholars can conduct audit studies, I want to suggest several ways that researchers can use this study type to examine social phenomena other than discrimination. One potentially promising direction would be to use email audit experiments to test the theoretical determinants of compliance. The idea here is that researchers can create experimental interventions that treat email responses (or non-responses, depending on the case) as a sign of participant adherence to some law, norm, or convention. For example, Terechshenko et al. (N.d.) investigated how international norms and the prospect of public sanctions might influence state respect for human rights. To examine the influence of these factors, they conducted an email audit experiment with a sample of 984 foreign diplomatic missions in the United States, Canada, and the United Kingdom. They emailed each mission

with a request for information about contacting domestic prisoners, a right that has long been acknowledged by the United Nations, and varied several attributes of the email. The important point is that receiving an email reply was interpreted as an act of compliance with an international norm. While a design like this could be extended to study compliance with other laws or norms, the main idea here is that email audit studies can be adapted to answer a wide range of substantive inquiries. Scholars have recently used email audit studies to examine the efficacy of economic regulations (Findley, Nielson and Sharman, 2015) among other phenomena. The only real constraint is the imagination of the researcher.

Another way of adapting email audit studies is to use them as the second part of a larger experimental design. For example, Butler and Crabtree (2017) conduct an experiment to reduce discrimination among public officials. In the first stage of their experiment, they sent a random sample of elected municipal officials an email that called attention to the growing literature on racial discrimination by political elites. In the second stage, they emailed nearly all elected municipal officials with requests for information, varying the racial identity of the putative constituent. They then examined whether the level of discrimination exhibited by officials in their treatment group was lower than the level of discrimination exhibited by officials in the control group.

This type of study suggests the potential of two-stage email audit studies. While Butler and Crabtree (2017) use this design to test the effect of an information treatment aimed at reducing bias, scholars can adapt this two-stage approach to examine the effect of other treatments on discrimination, compliance, and other types of sensitive behavior.

7 Discussion

In this chapter, I provided an overview of the steps involved from experimental design to empirical analysis. I then offered detailed recommendations about email address collection, email delivery, and email analysis, which are usually the three most challenging points of an audit study. The focus was on providing a set of primarily technical recommendations

to researchers who might want to conduct an email audit study. I concluded by suggesting several ways that email audit studies can be adapted to investigate a broader range of social phenomena. While going from the first to final stage in any email audit study can take considerable time, I think that the results they generate are often worth this cost. I hope that this chapter has helped reduce some of the effort for novice email auditors and thus encouraged the use of this simple but powerful study type.

References

Ahmed, Ali M, Lina Andersson and Mats Hammarstedt. 2012. "Does age matter for employability? A field experiment on ageism in the Swedish labour market." *Applied Economics Letters* 19(4):403–406.

Ahmed, Ali M, Lina Andersson and Mats Hammarstedt. 2013. "Are gay men and lesbians discriminated against in the hiring process?" *Southern Economic Journal* 79(3):565–585.

Baert, Stijn. 2016. "Wage subsidies and hiring chances for the disabled: some causal evidence." *The European Journal of Health Economics* 17(1):71–86.

Baert, Stijn, Jennifer Norga, Yannick Thuy and Marieke Van Hecke. 2016. "Getting grey hairs in the labour market. An alternative experiment on age discrimination." *Journal of Economic Psychology* 57:86–101.

Baert, Stijn, Sarah De Visschere, Koen Schoors, Désirée Vandenberghe and Eddy Omey. 2016. "First depressed, then discriminated against?" *Social Science & Medicine* 170:247–254.

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.

Broockman, David E. 2013. "Black politicians are more intrinsically motivated to advance Blacks' interests: A field experiment manipulating political incentives." *American Journal of Political Science* 57(3):521–536.

Bushman, Brad J and Angelica M Bonacci. 2004. "You've got mail: Using e-mail to examine the effect of prejudiced attitudes on discrimination against Arabs." *Journal of Experimental Social Psychology* 40(6):753–759.

Butler, Daniel M. 2014. *Representing the Advantaged: How Politicians Reinforce Inequality*. Cambridge University Press.

Butler, Daniel M and Charles Crabtree. 2017. "Moving Beyond Measurement: Adapting Audit Studies to Test Bias-Reducing Interventions." *Journal of Experimental Political Science* .

Butler, Daniel M and David E Broockman. 2011. "Do politicians racially discriminate against constituents? A field experiment on state legislators." *American Journal of Political Science* 55(3):463–477.

Coffman, Lucas C and Muriel Niederle. 2015. "Pre-analysis plans have limited upside, especially where replications are feasible." *The Journal of Economic Perspectives* 29(3):81–97.

Costa, Mia. N.d. "How Responsive are Political Elites? A Meta-Analysis of Experiments on Public Officials." . Forthcoming.

Crabtree, Charles, Matt Golder, Thomas Gschwend and Indriði H. Indriðason. N.d. Campaign Sentiment in European Party Manifestos. Technical report Working Paper.

Desposato, Scott. 2015. *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*. Routledge.

Driscoll, Jesse. 2015. "Prison States & Games of Chicken." *Ethics in Comparative Politics Experiments* .

Findley, Michael G, Daniel L Nielson and JC Sharman. 2015. "Causes of noncompliance with international law: A field experiment on anonymous incorporation." *American Journal of Political Science* 59(1):146–161.

Franco, Annie, Neil Malhotra and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science* 345(6203):1502–1505.

Fujii, Lee Ann. 2012. "Research ethics 101: Dilemmas and responsibilities." *PS: Political Science & Politics* 45(04):717–723.

Gaddis, S Michael. 2014. "Discrimination in the credential society: An audit study of race and college selectivity in the labor market." *Social Forces* p. sou111.

Gaddis, S. Michael. 2017*a*. The Ethics of Audit Studies: Legal Precedent, Best Practices, and Academic Cooperation. In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, ed. S. Michael Gaddis. Methodos Series New York, NY: Springer.

Gaddis, S. Michael. 2017*b*. An Introduction to Audit Studies in the Social Sciences. In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, ed. S. Michael Gaddis. Methodos Series New York, NY: Springer.

Gaddis, S Michael and Raj Ghoshal. 2015. "Arab American Housing Discrimination, Ethnic Competition, and the Contact Hypothesis." *The ANNALS of the American Academy of Political and Social Science* 660(1):282–299.

Gelman, Andrew and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gerber, Alan S and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation*. WW Norton.

Giulietti, Corrado, Mirco Tonin and Michael Vlassopoulos. 2015. "Racial Discrimination in Local Public Services: A Field Experiment in the US.".

Grose, Christian R. 2014. "Field Experimental Work on Political Institutions." *Annual Review of Political Science* 17.

Hauck, Robert JP. 2008. "Protecting Human Research Participants, IRBs, and Political Science Redux: Editor's Introduction." *PS: Political Science & Politics* 41(03):475–476.

Heckman, James J. 1998. "Detecting discrimination." *The Journal of Economic Perspectives* 12(2):101–116.

Hogan, Bernie and Brent Berry. 2011. "Racial and ethnic biases in rental housing: An audit study of online apartment listings." *City & Community* 10(4):351–372.

Imai, Kosuke, Marc Ratkovic et al. 2013. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics* 7(1):443–470.

Lahey, Joanna N and Ryan A Beasley. 2009. "Computerizing audit studies." *Journal of economic behavior & organization* 70(3):508–514.

Lin, Winston and Donald P Green. 2015. "Standard Operating Procedures: A Safety Net for Pre-Analysis Plans." *Berkeley. Retrieved from www. stat. berkeley. edu/˜ winston/sop-safety-net. pdf (2014). Promoting transparency in social science research. Science (New York, NY)* 343(6166):30–1.

Lohr, Sharon. 2009. *Sampling: design and analysis*. Nelson Education.

Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)*. pp. 55–60.

Milkman, Katherine L, Modupe Akinola and Dolly Chugh. 2012. "Temporal Distance and Discrimination An Audit Study in Academia." *Psychological Science* 23(7):710–717.

Milkman, Katherine L, Modupe Akinola and Dolly Chugh. 2015. "What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations." *Journal of Applied Psychology* 100(6):1678.

Moore, Ryan T and Keith Schnakenberg. 2012. "blockTools: Blocking, assignment, and diagnosing interference in randomized experiments." *R package version 0.5-7.[ht tp://rtm. wustl. edu/software. blockTools. htm]* .

Neumark, David, Roy J Bank and Kyle D Van Nort. 1995. Sex discrimination in restaurant hiring: an audit study. Technical report National Bureau of Economic Research.

Oh, Sun Jung and John Yinger. 2015. "What Have We Learned From Paired Testing in Housing Markets?" *Cityscape* 17(3):15.

Olken, Benjamin A. 2015. "Promises and perils of pre-analysis plans." *The Journal of Economic Perspectives* 29(3):61–80.

Pager, Devah and Hana Shepherd. 2008. "The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets." *Annual review of sociology* 34:181.

Pedula, David S. 2017. Emerging Frontiers in Audit Study Research: Mechanisms, Generalizability, and Variation. In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, ed. S. Michael Gaddis. Methodos Series New York, NY: Springer.

Pennebaker, James W. 2015. "LIWC: How it works (http://liwc.wpengine.com/how-it-works/).".

Radicati, Sara and Quoc Hoang. 2011. "Email statistics report, 2011-2015." *Retrieved May* 25:2011.

Riach, Peter A and Judith Rich. 2002. "Field experiments of discrimination in the market place." *The economic journal* 112(483):F480–F518.

Riach, Peter A and Judith Rich. 2004. "Deceptive Field Experiments of Discrimination: Are They Ethical?" *Kyklos* 57(457-470).

Sharman, Jason C. 2010. "Shopping for anonymous shell companies: an audit study of anonymity and crime in the international financial system." *The Journal of Economic Perspectives* 24(4):127–140.

Stocksdale, Michael. 2013. “E-mail: Not dead, evolving.”.

Suresh, K. 2011. “An overview of randomization techniques: an unbiased assessment of outcome in clinical research.” *Journal of human reproductive sciences* 4(1):8.

Terechshenko, Zhanna, Charles Crabtree, Kristine Eck and Christopher J. Fariss. N.d. International Norms, Sanctioning, and Prisoners’ Rights: A Field Experiment with Foreign Missions. Technical report Working Paper.

Turner, Margery A, Stephen Ross, George C Galster and John Yinger. 2002. Discrimination in metropolitan housing markets: national results from phase 1 of the housing discrimination study (HDS). Technical report.

Yanow, Dvora and Peregrine Schwartz-Shea. 2008. “Reforming institutional review board policy: Issues in implementation and field research.” *PS: Political Science & Politics* 41(03):483–494.