CHAPTER 3

Audit Studies in Political Science*

Daniel M. Butler and Charles Crabtree

Abstract

Audit studies typically involve researchers sending a message to or making a request of some sample in order to unobtrusively measure subjects' behaviors. These studies are often conducted as a way of measuring bias or discrimination. We introduce readers to audit studies, describe their basic design features, and then provide advice on effectively implementing these studies. In particular, we provide several suggestions aimed at improving the internal, ecological, and external validity of audit study findings.

3.1 Introduction

Audit studies[1] are part of a larger group of field experiments designed to measure, but not necessarily change, the behavior of subjects in the field.[2] The usefulness of these studies depends on whether they measure behavior in an unobtrusive manner (more on this later in the chapter). At their most basic design level, these studies

study discrimination in housing and labor markets (see review in Quillian et al. 2017). The civil rights movement led to the passage of legislation barring discrimination, which was accompanied by an interest in measuring whether discrimination persisted (see discussion in Gaddis 2017). Many of these studies looked at whether racial minorities were treated worse than their White counterparts (Wienk et al. 1979). Because audit studies on racial discrimination have been conducted for many decades, researchers have been able to compare how the treatment of racial minorities has changed (or not changed) over time (e.g., Quillian et al. 2017). While audit studies continue to study the treatment of racial minorities relative to Whites, the approach has also been applied to understand discrimination against many other groups. It has been used to examine discrimination based on one's gender (Ayres and Siegelman 1995; Butler 2014), age (Ahmed et al. 2012), sexual orientation (Drydakis 2014), religion (Adida et al. 2010; Pfaff et al. 2019), criminal record (Pager 2003), and other attributes (Gell-Redman et al. 2017; Rivera and Tilcsik 2016; Weichselbaumer 2016).

Audit studies have also been widely used by governments as a way to test for discrimination. In the 1960s, the UK parliament created the Race Relations Board, which commissioned several studies, including audit studies, aimed at measuring levels of racial discrimination (Daniel 1968). The tests uncovered discrimination and led to the passage of more laws barring racial discrimination in housing and employment (Smith 2015). In the USA, the US Department of Housing and Urban Development (HUD) conducted several audit studies to measure levels of discrimination in the housing market. In addition to several studies that focused on specific cities (e.g., Johnson et al. 1971; Pearce 1979), HUD commissioned several national audit studies (Turner and James 2015; Wienk et al. 1979; Yinger 1991, 1993). The federal government's decision to use audit studies to measure discrimination influenced academics by signaling that these studies were reliable, effective ways of measuring discrimination (see discussion in Gaddis 2017).

The advantage of audit studies, and all field measurement studies, is that they provide researchers with a measure of how the subjects under study behave. Survey responses can be

cheap talk, especially when the topic under investigation involves behavior that is socially unacceptable. Findley and Nielson (2016), for example, conducted follow-up surveys with some of the same companies who had been part of a field measurement study they had done looking at levels of compliance with international law. In the original study (Findley et al. 2015), the authors studied whether the individuals who provide incorporation services (i.e., helped with legal documents to create a company) for anonymous shell companies (i.e., companies that do not have employees but hold financial resources) are abiding by the international agreements related to the formation of these companies. To do so, they posed as consultants seeking to form anonymous shell companies and approached thousands of individuals who provide incorporation, varying where the citizen making the offer came from and whether a premium was offered for confidentially. They found that a large number of providers are willing to provide the service without the required identification documentation. In the follow-up survey, they asked respondents what documentation they would require if someone asked for help in creating an anonymous shell company. They then compared the survey responses to the behavior they observed in their field measurement study.

The results of their study show that the survey results overstate the level of compliance with the law. There are two reasons why the survey results understated the level of bad behavior. First, some of the worst offenders did not complete the survey. Second, those who responded to the survey self-reported higher levels of compliance with the law than what was observed in practice during the audit portion of the study (see also Doherty and Adler 2020; Pager and Quillian 2005).

Similar factors are likely to be a concern in any survey on discrimination. First, the people who are most discriminatory may be more likely to opt out because they know that their behavior is wrong or sense that it will be judged as wrong. If we try to draw conclusions based on the people who opt in, we might underestimate the level of bias. Audit studies can mitigate this issue by getting responses from a larger set of the sample of interest.

Second, social desirability is probably an issue in these contexts. Discrimination – the focus of

many audit studies – is the type of topic that is likely to suffer from social desirability effects. People may simply underreport their own discriminatory behaviors and attitudes in a survey. Audit studies avoid this potential pitfall by looking at individual behavior when people do not realize they are being studied (and thus cannot artificially change their behavior to look less biased to the researcher). The audit study, if well done, captures behavior in action, unaffected by social desirability bias.

3.2 Basics of an Audit Study

Most audit studies involve testing for discrimination in how a group responds to some type of request (an email seeking help, a job application, a housing application, etc.). Audit study designs generally involve the following steps:

- Identify the question, the population, and the sample.
- Develop the instrument(s).
- Randomly assign treatments
- Send messages.
- Measure the outcome by looking at responses.

To illustrate these steps, consider studying whether bureaucratic offices exhibit racial discrimination against Blacks relative to Whites. Contacting a government office with a request for help is just one way that audit studies can be applied. Other applications, as mentioned, include applying for jobs or housing or trying to complete other important quotidian tasks (see review in Gaddis 2017). We use the example of contacting a bureaucratic office because this is representative of an approach commonly used in political science (also see Chapter 27 in this volume). Researchers could study this question by sending email requests for help to different bureaucratic offices and randomizing whether the request comes from someone who is putatively Black or White (e.g., using stereotypical names to signify race). The requests would be identical in all ways other than the race of the requester. The researchers can then measure how the offices respond to the requests to see if they are less (or more) responsive to requests from Blacks.

By following these steps, researchers can measure levels of discrimination – if the response rates differ on average, then it suggests

discriminatory behavior based on race. In the rest of this section, we highlight a few of the major decisions that go into conducting an audit study.

3.2.1 Be Precise about the Question and the Population and Sample

Audit studies are well suited for studying discrimination. We follow Pager and Shepherd (2008) and define discrimination as the difference between how two groups are treated. Discrimination, which involves behavior, is distinct from holding racist attitudes or beliefs (e.g., prejudice). All of these other factors can motivate behavior. Studying discrimination does not presume what is causing the unequal behavior (see discussion in Pager and Shepherd 2008), though a promising, necessary direction of future work would be to examine potential causes.

Most audit studies, though not all (e.g., Butler et al. 2012), have focused on measuring whether subjects are engaging in some form of discrimination. In our running example, the researchers are interested in testing whether bureaucratic offices are discriminating based are race (e.g., are less responsive to Blacks than Whites). An audit study is appropriate for this question because it is focused on the behavior of government workers.

It is also important to be precise about the population and sample. In many existing audit studies, researchers contact legislative offices (Butler and Broockman 2011; Gell-Redman et al. 2018). Even if the researchers use the legislator's email address, these requests may be dealt with by staff. In other words, these studies are not necessarily about legislators, but rather about legislative offices. These studies speak to the behavior of legislative offices, which can tell us something about how legislators represent their constituents (Salisbury 1981a, 1981b). This is not to say that these studies are not informative; rather, it is important to be precise about who or what we are studying.

A related issue is whether the researcher is interested in the behavior of all legislators in the USA or just a specific subset (e.g., state legislators). Often researchers are interested in the behavior of all legislators but choose only to include state legislators in their sample. In many cases, this will be appropriate because all

legislators face similar electoral and party pressures (see discussion in Butler and Powell 2014). However, the researcher needs to evaluate this decision on a case-by-case basis: Is it appropriate to generalize from the sample to a larger population? The researcher should be clear about the population that they want to study and ensure that their sample is correct for the question of interest. In our running example, the researchers are interested in learning about how bureaucratic offices deal with requests.

3.2.2 Develop the Instrument (or Message) to Maximize the Likelihood That It Reflects a Commonly Encountered Communication

Audit studies are typically used to measure the level of discrimination that individuals face. This is best done by creating an instrument (or message) that people are likely to send. "Instrument" refers to the message that the researcher is sending to the people being studied. In a study of job market discrimination, the instrument might be the resume used to apply for jobs. In our running example, the instrument would be the email message that the researchers send to the bureaucratic offices.

It is crucial that the researchers develop the instrument so that the people in the sample being studied approach the communication as they normally would. If the people in the sample suspect that they are being studied, they may behave differently, leading the researcher to make incorrect conclusions. This point is so important that we devote a full section below (Section 3.3.1) to discussing it.

Finally, note that an instrument can vary in more than one aspect. Bertrand and Mullainathan (2004), for example, vary both race and qualifications in their experiment looking at how race influences employers' interest in job applicants. They find that the effect of race varies with the applicant's qualifications (the bias is worse for higher levels of qualifications). Researchers with theoretical reasons to vary more than one aspect can do so using a factorial design.

3.2.3 Randomly Assign Treatments

One decision that researchers have to make is

whether they will send just one message to each unit in the sample or multiple messages. Sometimes researchers will use a paired design, where they send each unit in the sample one message for each of the treatments. In our running example, a researcher using a paired design would send two (or more) messages to each bureaucratic office: one from a putatively Black individual and one from a putatively White individual. While this design can be appropriate and can increase statistical power, we generally recommend against it. As we mentioned above, audit studies are most effective when they unobtrusively measure respondents' behavior. A paired design increases the likelihood that the experiment may be discovered, which, as we discuss later in the chapter, hurts and potentially fatally compromises the usefulness of the study. In our running example, the researcher would randomly assign each office to receive a message from either a putatively Black individual or a putatively White individual.

3.2.4 Sending Messages That Hold Confounding Factors Constant

Early audit studies involved having actors from different racial groups apply for jobs or housing (e.g., Pager 2003). The actors would apply in person and the researchers would see whether the racial minorities were treated differently. One concern about these studies is that the White and racial minority actors likely differ in systematic ways. If these differences are also relevant to the hiring or housing decision, then it is possible that these confounding factors might be responsible for the differential treatment.

To avoid this potential criticism, researchers would identify people who were similar to begin with, and they would train auditors to respond in similar ways. The goal was to minimize any potential confounding characteristics. However, because it is nearly impossible to deal with all potential con-founders, including the fact that the auditors (i.e., the actors) knew the treatment, skeptics raised concerns that any measures of bias were inaccurate (e.g., even if their resumes were virtually identical, their in-person behaviors likely differed) (Heckman 1998).

In response to these criticisms, researchers have transitioned to sending messages by mail or email.[3] This allows them to send messages that are identical except in ways that researchers

intentionally manipulate. Returning to our running example, researchers might send email messages to bureaucratic offices that are the same in every way except in the name of the sender, which signals the sender's race, gender, or other ascriptive attributes.

3.2.5 Measuring Outcomes

Audit studies have a relatively clear interpretation, which makes them an attractive tool for studying how officials treat individuals from various groups.[4] Many of the audit studies in political science have been used to compare how public officials treat different groups of individuals (see the review in Costa 2017). The most common outcome is to look at whether a response is given to the sender's request or application. People sending the messages generally want a response, and so researchers can look at whether they receive one. This is generally a good outcome to report when performing an audit study, as it is the most basic outcome and can be compared with previous audit studies.

Researchers can also look at the quality of the response (the length of the response, the friendliness, whether the requested information was provided, etc.). Researchers who look at these outcomes must be careful to avoid the bias that comes from conditioning on whether the original message received a response (Hemker and Rink 2017). Whether a response was given is a post-treatment variable because it comes after the individual being studied has been exposed to their assigned treatment. For example, it might be that officials who are hurried for time are willing to provide a quick response to an email from a White individual but not a Black individual. Perhaps these quick responses would only be of medium quality, because the official was rushed, but they would still be better than no response. If, in this hypothetical case, the researcher limited the sample to people who responded, they could easily conclude that the White individuals received lower-quality responses because these quick responses bring the average quality down. While this example is hypothetical, the more general point is that conditioning on a post-treatment variable can introduce serious bias in unknown directions and should thus be avoided (Montgomery et al. 2018).

Coppock (2019) outlines three options for identifying unbiased estimates related to the

content of the email. Here, we discuss the approach that we believe will be best for most applications: redefining the outcome to avoid conditioning on whether a response was received. Redefining the outcome is straightforward and results in a dependent variable that is easy to interpret. When taking this approach, the researcher would redefine the outcome to be an indicator variable that is coded as 1 if it meets some outcome and 0 otherwise. Returning to our running example, it might be the case that the researcher has coded whether a response from the bureaucratic office answered the question. The redefined outcome could be: Does the office send a response that answered the question? It would be coded as 1 if the response answered the question and 0 if the response did not answer the question or no response was sent.

3.3 Maximizing Internal, Ecological, and External Validity

Researchers should aim to maximize the ecological validity of their study, or the extent to which it approximates real-world interactions. They can help achieve this goal by ensuring the realism of their instrument. Specifically, they should design an instrument that avoids raising study subjects' suspicions. If the study population suspects that the message they receive is not typical, they might doubt the identity of the sender or the purpose of the communication. For instance, if a law enforcement agency received a request about becoming a police officer from an individual that describes themselves as "Black" three times in an email, the individual reading the email might consider this unusual behavior and suspect that the message was testing how they would reply to a putatively Black citizen.

These doubts could change how they respond, potentially biasing the results away from the very thing the researchers want to learn from a study like this. For example, perhaps subjects normally responds more to Whites than Blacks. However, if they suspect that researchers are studying their behavior, they might be more careful in responding to communications from Blacks, to the point where they respond more to Blacks than Whites. This could lead researchers to mistakenly conclude that Blacks receive better, not worse treatment.[5]

Researchers should also maximize the realism

of their instrument related to what they want to learn from their audit studies. Typically, researchers conduct audit studies because they want to learn about how the average member of a group is treated in some interaction. If the study population suspects that the message that they receive is not typical, they might think that the individual sending it is also atypical in some way. This could cause them to treat the sender differently from the average member of the sender's group. For example, a researcher pretending to be a parent might send emails to a sample of principals that are unusually impolite. The principals might infer from the language used in the emails that the putative parent is entitled, bossy, or possesses some other negative personality attribute(s). As a result, they might reply differently from how they would to a message that was more polite and deferential.

3.3.1 Include Typical Requests in the Instrument

One aspect that affects the realism of the instrument is the type of request(s) included in the instrument. Before determining what request(s) to make of the study population, researchers should ensure that these request(s) are similar to the ones that their subjects usually receive. They can do this in several ways. One approach is to conduct qualitative interviews with members of the study population about the type of interactions that they have with the public (Terechshenko et al. 2019).[6] Researchers conducting audit studies to examine how offices (legislative, bureaucratic, etc.) behave might also use requests that appear in the frequently asked questions sections of office websites. When the study population consists of public offices or officials but neither of these two approaches is possible, researchers might want to consider issuing Freedom of Information Act (FOIA) requests for all messages received by the study population.[7] After receiving these text corpora, researchers could use machine learning tools to summarize them and to construct typical requests (Grimmer and Stewart 2013).

3.3.2 Use Different Aliases

In the majority of audit studies, researchers create a set of fictitious identities and use these to send messages. Sometimes researchers use these identities to detect discrimination, sometimes

they use them to conceal the fact that the messages come from researchers. The names that researchers use with these identities should be carefully selected. This is because each name signals a number of things about the identity of the sender. In the interests of maximizing the believability of the messages, researchers should select names that are typical among members of the individual identity's group. When researchers do this, they should check how the names they use are perceived because perceptions of names vary across study populations based on subject race, education, and geography (e.g., Crabtree and Chykina 2018). Researchers have two options for dealing with this potential problem. One is that they can use names that have been tested by other researchers. For example, Hughes et al. (2019) provide in their appendix a list of the names that they used, along with the results of a survey they conducted about popular ethnic perceptions of these names. Researchers can also pretest their own names through platforms such as Amazon's Mechnical Turk. In the context of our running example, one could do this by selecting a large bundle of names and asking Mechnical Turk workers (MTurkers) to assess the likelihood that the name belongs to a Black or White individual.[8] The results from this exercise would allow researchers to select the names that are most strongly associated with Black or White individuals.

Regardless of the approach researchers use to expand their battery of names, we suggest that they pick names that are similarly perceived. Unfortunately, there are multiple ways to measure similarity, and there are no hard rules about what counts as "similar" enough. Researchers should rely on their own contextual, theoretical, substantive knowledge when making this choice and transparently report their decision rule in their description of the design.[9]

3.3.3 Use Multiple Requests and Names

A potential issue for researchers is that one of their subjects might receive multiple messages or that their subjects might share received messages with each other. This is potentially problematic if those messages are identical in nearly all aspects. That might lead subjects to doubt the authenticity of the messages or discern the intentions of the

researchers, effectively spoiling the experiment. One way in which researchers can potentially get around this issue is by randomly varying aspects of the instrument, such as the included requests and sender names. In our running example, we might send several different types of requests to bureaucratic offices and use several different names to signal Black and White identities. By doing this, we would make it less likely that the messages would seem related to each other, which decreases the chances of potential discovery. As above, we suggest that researchers leverage their understanding of the phenomena they are studying to determine how similar the requests need to be.

There is a second compelling reason to use a range of requests and names. Researchers often want to claim that the results of their study are indicative of more general social phenomena. By using different requests and different names, researchers can ensure that their results are not specific to any one request or name. In doing so, they can help maximize the external validity of their study.

3.3.4 Use Reasonable Email or Postal Services

Once researchers have developed an instrument, they need to deliver it. To do that, researchers typically create email or mail addresses for each identity that they use in their study. To maximize the believability of their intervention, researchers should use addresses that do not raise subject suspicions. This can be done by using common email or mail services, such as gmail.com or a post office box. If researchers are using email to deliver their instrument, they should consider creating unremarkable email addresses. For example, if the name for one identity is "Jane Smith," they might want to create the email address "jsmith872998@gmail.com."[10] In some cases, researchers might want to have their identities associated with a real or fictional organization. This might lead them to partner with organizations and use their email domains to increase the believability of their messages.

3.3.5 Check the Final Instrument

Once researchers have a draft of the final instrument, they should perform two additional checks. First, they should ensure that their

instrument is not the same as one used in a prior study. As a corollary, they should not borrow parts of their instrument from previously completed studies. This can cause significant problems for researchers. As an example, White et al. (2015) used a set of names and email addresses to determine whether local election officials exhibited bias against Latinos in September 2012. Approximately four years later, a researcher used the same set of identities in an unrelated project. Some local election officials detected the similarities and posted a notice on a public bulletin board that individuals should not respond to these emails, in effect contaminating the researcher's study (Kovaleski 2019). Second, researchers should ask several individuals who are or have been part of the study population to read the instrument and provide input (Pfaff et al. 2019).[11] In our running example, we could ask individuals who used to work at bureaucratic offices what they thought about our instrument. These exchanges between researchers and the subject population can help identify issues with the instrument or suggest new ways of improving it or the broader experimental design.

3.4 Additional Design Considerations

3.4.1 Internal Validity

As we discuss above, the internal validity of any empirical claims made from audit studies depends on the subject pool not knowing that they are the participants in an experiment. Just as importantly, the internal validity of these claims also depends on the identities used by the auditors appearing identical (in both observed and unobserved ways) to participants, with the exception of whatever attributes researchers intentionally manipulate (Heckman 1998). When identities do not otherwise appear identical, then researchers cannot be sure that any discrimination that they measure is related to the characteristics that they manipulate or to some other characteristics that correlate with them and might vary across identities. Another way of thinking about this is that the results from audit studies depend on the excludability assumption that the manipulated characteristic drives differences in how subjects respond and not some other characteristic (Butler and Homola 2017;

Gerber and Green 2012).

This concern is part of the reason why most audit studies are conducted via correspondence now, rather than in person (Gaddis 2017; Neumark 2012). For example, by sending messages to subjects, researchers have more control over how they construct identities, making it theoretically easier to create similar messenger profiles. In our running example, it would be more feasible to create two constituents who share all visible characteristics except for their race, as signaled via their name, than to find a White person who is interchangeable with a Black person in every other way except for their race.

Even in this case, though, people might object that the excludability restriction does not hold, and that the names that researchers use might signal not only race, but other characteristics as well (Fryer and Levitt 2004). For example, one might object that a name like "Lakisha" not only indicates that the putative sender might be Black, but also that they come from a poorer or less educated family (Butler and Homola 2017; Gaddis 2017). Thankfully, there are a variety of ways in which we can empirically assess the extent to which identities might appear the same. One approach is to pretest the different identities with some survey population, such as MTurkers (Gaddis 2018). The idea here is to ask respondents a series of questions about each identity's observed and unobserved characteristics.[12] If the responses indicate that the only differences relate to the manipulated characteristics, then researchers can be more confident that they have not failed to set some attributes as constant. On the other hand, if the responses indicate that names are different across multiple dimensions – "Archibald" and "Jamal" likely signal both race and socioeconomic status – then researchers can adjust their lists of names accordingly.

3.4.2 Spam Concerns

One potential concern with conducting audit studies via email is that messages might be automatically marked as Spam. This would mean that subjects might not receive their assigned treatments. This would potentially be very problematic if certain experimental treatments or treatment combinations were more likely to be identified as Spam. The issue here is that this

would decrease the probability that messages with those treatments would receive a reply, potentially leading researchers to believe that bias exists where it does not.

To help