NBER WORKING PAPER SERIES

TECHNICAL ASPECTS OF CORRESPONDENCE STUDIES

Joanna Lahey
Ryan Beasley

Working Paper 22818
http://www.nber.org/papers/w22818

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2016

I have no disclosures to make. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Technical Aspects of Correspondence Studies
Joanna Lahey and Ryan Beasley
NBER Working Paper No. 22818
November 2016
JEL No. C9,C93,J2,J7

ABSTRACT

This paper discusses technical concerns and choices that arise when crafting a correspondence or audit study using external validity as a motivating framework. We will discuss resume creation, including power analysis, choice of inputs, pros and cons of matching pairs, solutions to the limited template problem, and ensuring that instruments indicate what the experimenters want them to indicate. Further topics about implementation include when and for how long to field a study, deciding on a participant pool, and whether or not to use replacement from the participant pool. More technical topics include matching outcomes to inputs, data storage, and analysis issues such as when to use clustering, when not to use fixed effects, and how to measure heterogeneous and interactive effects. We end with a technical checklist that experimenters can utilize prior to fielding a correspondence study.

Joanna Lahey
The Bush School
Texas A&M University
Mailstop 4220
College Station, TX 77843
and NBER
jlahey@nber.org

Ryan Beasley
954 Melvin Rd
Annapolis, MD 21403
dr.ryan.a.beasley@gmail.com

Technical Aspects of Correspondence Studies

Joanna Lahey and Ryan Beasley

Abstract This chapter discusses technical concerns and choices that arise when crafting a correspondence or audit study using external validity as a motivating framework. The chapter discusses resume creation, including power analysis, choice of inputs, pros and cons of matching pairs, solutions to the limited template problem, and ensuring that instruments indicate what the experimenters want them to indicate. Further topics about implementation include when and for how long to field a study, deciding on a participant pool, and whether or not to use replacement from the participant pool. More technical topics include matching outcomes to inputs, data storage, and analysis issues such as when to use clustering, when not to use fixed effects, and how to measure heterogeneous and interactive effects. The chapter ends with a technical checklist that experimenters can utilize prior to fielding a correspondence study.

1 External Validity and the Audit Study

External validity concerns drive many technical choices in correspondence studies. While it is tempting to believe that a single study can answer "Is there *X* discrimination?" or "Do for profit colleges and universities provide value?", an audit study can only test a limited market for a specific subset of applicants during a specific time period. It is therefore vital to design the experiment carefully, so that it is clear how the study's results will further knowledge. In general we will use examples from employment audit studies to illustrate ideas in this chapter, but correspondence review is a powerful tool that can be used more broadly to study differential treatment across many settings.

Ultimately, the external validity of an experiment is constrained by each decision made in the design. For example, studies that only apply to ads within big cities may not be applicable to smaller towns or rural areas. Similarly, resumes in which every person over the age of 50 also has a multi-year employment gap may provide results that are driven by the age, by the gap, or by their combination. Questions to ask in the initial design phase include: Who will you use as

J. Lahey, Texas A&M University and NBER, email: jlahey@tamu.edu
R. Beasley, SimQuest Solutions Inc.
This draft has been prepared for the volume *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, edited by S. Michael Gaddis. The authors thank all of the researchers who have provided feedback on the Resume Randomizer program, and Joanna Lahey also thanks the many editors who, through referee requests, have forced her to keep up-to-date on the state of correspondence studies. Thanks also to Patrick Button for helpful feedback.

participants? When and for how long will you field the study? Where will you get correspondence inputs? Taking the design as a whole, for what group will the results of the experiment be externally valid?

The most important external validity question to ask is whether the indicator that separates the treatment group(s) from the control group tests what it is supposed to test and does not inadvertently test something different. Examples of indicators include names for race discrimination (e.g. Bertrand and Mullainathan 2004; Gaddis ch. 8 this volume, Oreopoulos 2011), date of school graduation for age (e.g. Lahey 2008, Neumark, Burn, and Button 2016), or name of school when testing the effect of for-profit colleges (e.g. Gaddis 2014; Deming et al. 2016; Darolia et al. 2015). It is important that the indicator indicates what it is intended to indicate and is not just measuring that a resume or other piece of correspondence is unusual. For example, indicating age by date of high school graduation is something that most real job seekers do, but listing age on a resume is frowned upon in the United States. The most troubling examples occur when the unintended "unusual" negative signal only signals negatively for treatment. For example, putting union membership on a nursing resume is not just testing the effect of union status, and similarly, listing number of children does not just indicate that the applicant is a mother, but that the applicant does not know not to put things on a resume that do not belong there. It is our belief that if this indicator is "unusual" rather than something that normally appears in resumes that the study should not be performed and resources should be devoted elsewhere. As a caveat, avoiding testing "unusual" items does not mean that it would be inappropriate for someone from a discriminated group to apply for a position. Men do apply for clerical jobs, women do apply for truck driving positions, and minorities do apply for high powered jobs; the general equilibrium employment ratio does not necessarily indicate that applicants are not interested in a job.

A final external validity concern is the open question of how call-backs translate into job offers. It is important to note that this translation will be different for different types of jobs. Although there are scattered answers to this question from various industry surveys and studies (e.g. Barron et al. 1985, Howden 2016, Maurer 2016) and studies of job seekers (e.g. Moynihan et al. 2003), we do not know what the average translation from call-back to job offer is or how this number varies by industry, occupation, unemployment rates, educational level of the applicant and so on. During the design phase, it is important to investigate how actual job-seekers enter the selection process and to be careful about making broader claims on how interviews translate into hires.

While decisions driven by external validity motivations should guide study design, this chapter will also discuss technical considerations including power analysis, matching outcomes to inputs, data storage, how to deal with changes while fielding the experiment, and post-collection data analysis concerns. The chapter ends with a technical checklist to aid researchers.

2 Determining the Pool

2.1 Participants

In a correspondence audit, the participants will generally be companies, landlords, purchasers, and so on, that is, members of the group whose biases are being tested, not the hypothetical applicants. Results will only be externally valid to the pool of participants tested in the experiment. Results may be different if participants are drawn from, for example, urban vs. rural areas, from the Southern US region vs. the Northeast, or from Belgium compared to Mexico. Studies that cover a broad geographical area may be affected by heterogeneous effects across different cities or states

or countries, and will need to have a large enough sample size to be able to detect, and preferably test, those differences. On the other hand, because effects may be different across regions, results for one area are only externally valid for that area, thus broader geographic coverage may give effects that are externally valid on average even if they do not provide a good representation for what any individual faces in a smaller market.

The choice of how to find participants is important, especially in these times of rapid technological change. For example, in the past, classified ads in the newspaper were a primary way that jobs were posted, which meant that early studies could use Sunday want-ads in order to run an experiment that was externally valid for a large population of job seekers. Companies still take out classified ads in trade magazines/journals in order to reach a specific audience, but online resources have risen in prominence. Craigslist in particular is a popular website for researchers and has increased its market penetration across the United States. Online job sites such as Indeed.com or snagajob.com have also become more prevalent and potentially more useful than their earlier incarnations ten or twenty years ago.[1] Not all sites have the same job ad penetration across geographic markets or fields and a researcher should investigate these differences before committing to a specific source of advertising. What may be a good source of jobs for computer science positions may not be as good a source for nursing positions. Other researchers may avoid general want-ad postings and pick specific companies to target with unsolicited applications. This method could include, for example, targeting Fortune 500 companies (e.g. Bendick et al. 1999) or all of the hospitals and nursing homes in a specific area. In some regions researchers can use job banks, such as Belgium's job bank (e.g. Baert and Dieter 2016). Some professions rely on walk-ins or networking for the majority of their job openings and as such are more difficult to test in a correspondence framework (Holzer 1996). Again, these decisions should be guided by both feasibility and external validity of the sample to the question you are interested in answering. Carbonaro and Schwarz in Chapter 5 of this volume will go into more detail on these concerns.

When choosing specific participants, it is important to have a systematic rule in place that provides the most externally valid sample possible. For example, a simple rule could be to apply to all ads posted on your online site during the course of the study, checking for new ads once a day. Drawing a sample may require more complicated rules that should be decided on in advance or during a pilot study.

2.2 Length of time

Another important choice is when and for how long to field a survey. Using employment audits as an example, it is important to think about how business cycles might affect hiring. Results during the holiday hiring season, when many lower-level companies and job seekers are looking for holiday work, may be different than results during a hiring lull. A study that is externally valid for college students looking for summer work will not be externally valid for applicants searching for a full-time career. Similarly, many companies will advertise for positions after the first of the year, and industries tied to fair weather will advertise in the Spring (see JOLTS 2016 for data on hiring seasonality).

An additional factor that may determine how long you field a study is more practical—the necessary number of observations to find statistical significance for a given power, something we

[1] Note that, as always, you should check with your IRB about what job sites are allowable based on their Terms of Service (TOS). Some IRB allow TOS violations that could happen in the normal course of use, whereas others do not allow such usage.

discuss below in the section on power analysis. Similarly, the expected response rate can mechanically affect a study's ability to obtain variation even with a large sample size. Response rates can depend on the type of participants, the number of participants, whether the participants are actively hiring, annual cycles, long-term recessions or expansions/recoveries, how many resumes are sent to each participant, and participant strategies of satisficing vs optimizing. Expanding on that last point, for positions that expect a lot of turnover the participant may use a satisficing rule and hire the first number of applicants that meet certain criteria, so there may be more call-backs overall and the timing of when resumes are sent may be important for detecting differential treatment. Other job openings, particularly those with more limited positions and longer tenure, may use an optimizing rule in order to get the best applicant possible, so there may be fewer call-backs and the quality of the resume will be important for finding differential treatment.

2.3 How many pieces of correspondence to send

The choice of how many resumes to send to a single participant at one time has tradeoffs. An obvious benefit of sending multiple resumes to a single firm is that it is an easy way to increase the number of total resumes sent. As with matched-pairs designs, this choice makes it easier to see how a single participant treats different types of resumes and can help to make a compelling argument for differential treatment that media reporters can easily understand. However, the choice to send multiple resumes to a firm comes with several potential drawbacks. One problem common to any within-subject design compared to between-subjects design is that inclusion of different treatments and controls can cause the participant to more directly compare these treatments to each other than he or she would if only viewing one treatment or control, thus decreasing levels of detected discrimination. These types of effects are seen in experiments generally (e.g. Charness et al. 2012; Tversky and Kahneman 1981) and there is some evidence of spillover effects of resumes within audit studies themselves (Phillips 2016). A related problem is that with more hypothetical resumes, the participant may change his or her priors about the underlying quality distribution and number of potential employees within the applicant pool. Thus any results from these studies will be externally valid to a different sample than reality. With a large enough number of resumes sent for a small number of interviews, there may also be mechanical effects—weak levels of discrimination will be magnified if, with a smaller number of applicants, equivalent resumes from both groups would receive an interview. Finally, there may be ethical concerns if the number of resumes is large enough to affect the hiring manager's practices; he or she may have trouble hiring if, for example, the opening has received a larger number of highly qualified applications than usual because of a large number of hypothetical applications. There is not one right answer for how many resumes to send to an open position. The benefits and disadvantages will vary by job type. In general, the disadvantages will be lower with openings that receive a larger number of applications than those which receive a smaller number. For example, sending four resumes of varying quality to a low level job during a recession for an opening that receives hundreds of resumes will probably still produce externally valid results and not harm the company, but sending four high quality resumes for a job that has a pool of maybe twenty qualified applicants (e.g. Horton forthcoming) can provide biased results and harm the company.

Another decision to make regarding the participant pool is whether to sample participants with or without replacement. For example, if an employer advertises a second time during the

sampling period, will it receive multiple sets of resumes from the study? External validity concerns would suggest considering if an actual seeker would apply for the same job or company again. This answer may depend on the time between reposting, and if it is for the same job that has already rejected the applicant or for a different job in the same company. Sampling with replacement has some downsides, however. If the resumes sent are from a quality pool that is sufficiently different from real job applications to the firm, then the study itself may be changing the employers' beliefs about the applicant pool which may have spillovers to the results. Another design concern may make this decision mechanically—if the sets of resumes are similar but not identical across items, for example, they use the same names and contact information but the other resume items vary, then a second set of applications to the same firm will be testing the effect of seeing resumes for what seems to be the same applicant but with at least one set of qualifications forged. Again, this concern ties back to the original guideline to not test "unusual."

3 Crafting correspondence

3.1 Choosing correspondence inputs

After selecting the participant pool, the next question to address is how to build correspondence inputs. In general, correspondence should be both realistic and externally valid to the pool tested. A common tactic in employment audit studies is to take inputs from real resumes gathered from online resume banks. These inputs are then either mixed and matched or modified slightly and used for a different employment pool so as to not negatively interfere with the job search of the applicants whose inputs were used. Care should be taken with this strategy; while it may be more externally valid than entirely fabricating inputs, it is still only externally valid for applications of the same quality or composition group from which the inputs came. In particular, the quality pools for resume banks may differ greatly. For example, resume audits from the early 2000s often used Americasjobbank.com, which was a government-run job bank program. The resumes in this bank were often low quality, e.g., full of typographical errors. Resumes that remained in the bank for longer periods of time tended to be of especially poor quality. More modern resume banks, for example, Indeed.com, seem to have higher quality examples on average. There is no guarantee that the composition of resumes on a resume bank site is equivalent to the composition of resumes that a posted advertisement will receive.

Quality of correspondence is additionally important for theoretical reasons. For example, with theories of variance-based statistical discrimination there is an interaction between quality of the resume and the treatment variable, with the dominant group preferred at higher quality levels, but the group for which there is less information preferred at lower quality levels. If the quality distribution of correspondence is small, the experiment may only be able to pick up a portion of this activity and may potentially give misleading results about the market as a whole. If the question being asked focuses on a specific quality segment of the market, the correspondence quality will be less of a problem because the pool is externally valid to the question being asked. An additional concern with quality levels is a mechanical one—if the quality of correspondence is too low, it may be difficult to get any positive responses from participants; treatment and control correspondence will have been treated the same, but that does not prove the lack of discrimination in the labor market and the results will not provide useful information on the impact of individual resume characteristics and their interactions.

As discussed in the first section, the choice of indicator that separates the treatment group(s) from the control groups is a key decision in the study design. Particularly, researchers

should avoid correspondence that stands out for reasons unrelated to the study. Otherwise the external validity is reduced because the results show how participants treat unusual correspondence rather than showing how they treat the variable of interest.

It is important to be aware of trends in whatever area of correspondence being tested. For example, styles change with regard to resumes and are not consistent across countries. Using recommendations for how to create a resume from 10 or 20 years ago may show that the applicant has not kept up with the times; in this case the results would only be externally valid for the group of applicants who submit old-fashioned resumes. Objective statements have fallen in and out of favor, various sections on the resume are given more or less weight, what type and how much previous experience to include varies, and so on. What is true at the time this volume is being written may be outdated in ten years. Prior to starting a study, determine what is "normal" for the study's specific area of interest. In the employment context, this can be done via viewing actual resumes submitted for a recent job opening, talking to HR representatives or hiring managers for positions similar to the type you are testing, and reading recent popular advice for job seekers.

3.2 Creating Correspondence

Once inputs have been gathered and the indicator has been chosen, those elements are combined to create the correspondence. Early studies based on matched-paired audits would often have a small number of correspondence templates, perhaps as many as eight, that they manually assigned names of different races or genders. This type of study is only externally valid for the types of people similar to those that the template represents, making it impossible to get a full view or even a large view of the labor market. In addition, without variation within the templates, it is difficult if not impossible to get a full picture of who within the broad group is being discriminated against, how they are being discriminated against, and why they are being discriminated against.

Our previous paper (Lahey and Beasley 2009) addressed these concerns and argued that three common problems with audit experiments were surmountable through automated random generation of correspondence. First, with limited numbers of templates, all items except the variable of interest are correlated within each pair of templates, so the results can only predict the outcomes and interaction effects for specific bundles of characteristics rather than individual characteristics. We discuss this concern, which we term "template bias," in more detail later in this section. Second, experimenter bias is exacerbated when humans are responsible for manually generating correspondence or matching templates to jobs, because the human may subconsciously deviate from random assignment. Third, early in-person matched-pairs audits were limited in scale and scope by expense, which necessitated small sample audit analysis.

With automated, random, generation of correspondence, the number of templates is no longer limited because each correspondence can have some probability to contain any given characteristic, robust pseudo-random number generators replace human action and thus avoid experimenter bias, and (given sufficient input material) generating large numbers of unique correspondence is quick and inexpensive. With enough responses, standard econometric techniques (OLS or Probit/Logit) can be used to test the impact of individual correspondence characteristics and their interactions with group differences on the outcome of interest. Additionally, with many templates or completely unique randomized correspondence, the researcher can allow the market to determine what the quality of a resume is rather than imposing one's own beliefs about what employers are looking for, something we discuss in the analysis section. At the same time, each additional variable may decrease the power of the study. In

general, we are in favor of large audit studies that are powered for main pre-specified hypotheses but that also allow for tests of secondary hypotheses that the study may not have enough power to test.

A simple approach to generating correspondence is via "mail merge", a thirty-year-old method in which a form letter has blanks that get filled from a list of text inputs, e.g., names and addresses (Friedman et al. 2013). The resulting correspondence outputs are generally nearly identical because the majority of the text is unchanged. While straightforward to use and supported by most word processors (current versions of Microsoft Word have a Mailings tab with a "Start Mail Merge" option), mail merge does no more than fill form letters by copying text from a list of inputs. The experimenter must take care in creating the list of text inputs to avoid experimenter bias, then create a dataset to link correspondence characteristics to outputs, then prepare different form letters (i.e., templates) if extensively different correspondence is desired. If different blanks in a form should relate (e.g., employment history is a function of bachelor's degree) then the experimenter must create the list of text inputs to contain that relationship. So while mail merge can fill a form letter with input text, the experimenter must manually generate the form letter and the inputs, and is saved only the effort of copy/pasting the latter into the former. Thereby mail merge solves the small sample problem because it assists in generating more correspondence quickly and easily, but it does not help with either limited templates or experimenter bias.

To help in the implementation of audit studies that surmount all three problems, we have developed a free open-source computer program named Resume Randomizer[2]. The program can create correspondence with a large number of experimenter-defined characteristics, and comes in two parts. The first part is an HTML-based user interface used to create templates. These templates can randomize inputs across the correspondence, including specifying the probability that an input will be included or the number of times an input will be included. For example, each correspondence may start with the same salutation, then have a random slot that selects between many unique first sentences for an objective statement, then have another random slot that has a twenty-five percent chance of outputting nothing and otherwise randomly chooses four different job history statements, and so on.

The second part of the program is an executable that uses a template file to generate multiple correspondence to be sent to the same participant. This part of the program allows for "matching" between correspondence so that either all the correspondence generated have the same characteristic for a given item, or so that none of them share characteristics for that item. The generated correspondence are plain-text, but various approaches can be used to add formatting, including generating the correspondence in TeX or HTML, or once the characteristics are chosen via Resume Randomizer then using mail merge to put those characteristics into Word documents (Oreopoulos 2011). Along with each correspondence, the second part of the program saves a "variable file" that, when combined with the input texts and template, contains all the information necessary to re-create the correspondence. This variable file can be imported into a statistical program, e.g., Stata, to analyze the impacts of characteristics.

With this program, researchers can generate correspondence sufficient for using standard econometric techniques to test the impact of individual correspondence characteristics and their interactions. Researchers can create a "template" that avoids template bias; each characteristic has some probability of being placed onto each resume or letter, so the impact of each characteristic (or group of characteristics) can be tested separately. The problem of experimenter bias can be

[2] Available at http://www.nber.org/data/ (under "Other"), at https://github.com/beaslera/resumerandomizer, or from the authors by request.

mitigated because the software composes the correspondence randomly, so an un-biased template will lead to un-biased correspondence, in aggregate. As with mail merge, this program substantially reduces the expense of generating additional correspondence, though the researcher must still provide sufficient input texts.

Since the initial release, we have revised the Resume Randomizer program for clarity, additional features, and ease of use. Random sections can now be configured to specify the exact percentage chance of choosing each potential result. Sections of the template can now be chosen based on the selection made in a previous random section, e.g., fraternity vs sorority membership at the end of a resume can naturally depend upon a random gender choice at the start of the resume. Text can be saved into variables defined on-the-fly in the template, and then recalled from those variables later in the template, e.g., randomly choose the name at the top of the letter and save the corresponding initials for use later in the letter. To ease analysis, the executable now automatically generates a codebook that maps the variables saved in the variable file to the text that gets placed in the correspondence. To simplify assembly of the input text snippets, templates can now import text files that solely contain such text items. We will continue to incorporate useful features as we get feedback from users.

3.3 Matched correspondence

An important choice is whether or not to use matched pairs in the audits. This study design essentially sends two resumes to the same firm that are identical except for the group characteristic of interest. Matched pairs were originally used for in-person audits because they dramatically increase power for small sample sizes. For studies that are necessarily small, matched pairs may still be the best design choice. However, there are drawbacks that come with matching pairs in audit studies. Using matched pairs is a within-subjects study design rather than a between-subjects design, which means that the same participant sees both the treatment(s) and the control (Charness et al. 2012). Even if participants do not realize that they are participating in an experiment, they are more likely to make a direct comparison between the treatment(s) and control which may change the effects of discrimination, most likely decreasing them by reducing implicit bias (e.g. Olian et al. 1988). A more ethical concern is that sending a participant matched sets of correspondence may be more likely to distort the participant's view of the labor market if they think that a specific type of hypothetical applicant is more heavily represented in the labor market pool than is actually true. Unmatched sets send a less focused signal and may be less likely to harm a participant's overall view of the market.

It is possible that the matched-pairs design may be better able to test for differences in situations in which some element of what is being tested can affect the general equilibrium applicant pool. For example, a hypothetical resume audit could find that firms that advertise as being AA/EEOC are less likely to interview hypothetical black workers than firms that do not advertise as being AA/EEOC. These AA/EEOC firms may still be less discriminatory if general equilibrium effects of having AA/EEOC advertising mean that more black applicants are applying to the firm (Kang et al. 2016).[3] From the standpoint of a single minority job seeker the reason for not getting called for an interview is less relevant, but from the standpoint of the labor market we would not be able to make the claim that firms with AA/EEOC are more discriminatory. The

[3] See Pager and Pedulla (2015) for more information on how perceived discrimination affects job application behavior.

black/white comparison within firms that advertise AA/EEOC is important, and matched pairs may be the best way of getting enough power to test for these effects. Chapter 6 by Mike Vuolo will discuss concerns about matched pair audits in more detail.

4 Sample size

An important part of the experimental design phase is figuring out the minimum sample size needed to find significant results for a reasonable effect size given a set power. Determining necessary sample size via power analysis requires information on effect size, desired significance level and desired power. Ideally the effect size will come from a pilot study. However, it is possible to get suggested effect sizes for field experiments from previously completed laboratory work or from related field studies. Psychologists have long been interested in many of the questions that other social scientists are just now testing in the field. In the absence of any prior related work, experimenters can use the default effect sizes of small, medium, or large based on beliefs about the size of the effect or based on the practical impact of an effect that is small, medium, or large. That is, if it is believed that a small effect size would be unimportant for the population in question, then it may be sufficient to gather a sample that could only capture a medium size effect. In general, one can choose standard levels for significance (0.05) and power (0.8), although these heuristics may be overly simplistic (Cohen 1977, 1992).

Power analysis has become easier in recent years given the availability of the program G*Power.[4] Current versions of G*Power can even determine sample size for matched pair studies. While G*Power is remarkable in many respects, as of this writing, it still lacks in two areas important to researchers planning audit studies. First, G*Power does not take into account clustering. If the study design includes sending multiple pieces of correspondence to the same participant, G*Power does not account for how power is affected by the loss in variation due to multiple samples per participant. To take into account the additional sample size needed because of the clustered design, sample size calculations from multi-level modeling for two levels can be used.

$$Sample\ Size_{final} = Sample\ Size_{G*Power} * (1 + (number\ of\ items\ per\ cluster - 1) * ICC)$$

The desired sample size, $Sample\ Size_{final}$, is calculated by multiplying the sample size (given by G*Power) that does not take into account clustering by a factor that takes into account both the number of items per cluster (ex. the number of resumes being sent to a firm) and the average inter-correlation between clusters (ICC). With a pilot study, the ICC can be determined using the *xtmixed* or *mixed* commands in Stata to determine standard deviations and applying the following formula:

$$ICC = \frac{\sigma^2_{cons}}{\sigma^2_{cons} + \sigma^2_{residual}}$$

[4] Stata's currently supported sample size calculator is *power*, but as of this writing has limited options compared to G*Power and thus is only recommended for simple designs, although its *nratio* option is useful for unbalanced designs.

where σ^2_{cons} is the standard deviation of the constant and $\sigma^2_{residual}$is the standard deviation of the residual. In the absence of a pilot study, default ICC range from 0.10 to 0.30 (Gulliford et al. 1999; Maas and Hox 2005).

A second drawback of G*Power is that how to test power for interactive effects is unclear— the "Linear Regression Model" options do not provide information on power to test the significance of an interacted coefficient, but test the effect of the interaction on the regression's R^2. Instead, G*Power's ANOVA framework can provide sample size analysis for interactive effects.

5 Datamining Concerns: Pre-registration and mid-experiment analysis

Pre-registering experimental plans has become more de rigueur in recent years. Grant proposals, which are often necessary to pay for experimental work, function in a similar way to pre-registration because they force researchers to outline their hypotheses and analysis plans a priori. Olken (2015) does an excellent job explaining the pros and cons of pre-analysis plans. Such plans remove problems of data-mining and remove the need for most robustness checks, but also limit exploration and are difficult to implement for tests of more complicated theories. Our general belief is that there are benefits to plan pre-registration but that one should not be dissuaded from doing exploratory secondary analysis in conjunction with or after completing the primary analysis. Correspondence review studies are large undertakings and are often our first glimpse at the hiring sides of various markets. One correspondence review cannot provide the definitive answer to any economic question and there is a place for exploratory work that informs future pre-planned studies.

How often to analyze the data while the study is being run is a related concern that has trade-offs with data-mining. In the ideal world, researchers would design the study, do a small pilot study to make sure everything was in working order and to get information for sample size calculations, and then they would run the experiment without looking at the results until it had completed. In the real world, however, mid-stream checks are important to make sure that the experiment is still running smoothly and is free from human error or unforeseen external shocks. While it may be tempting to use mid-stream checks to make major changes in the experiment based on results, doing so comes at the expense of data-mining concerns.

6 Technical Data Concerns

6.1 Sending Correspondence

How resumes are submitted has changed over the past few decades. In early studies it was standard to mail applications or to submit them by hand. Studies from 15 years ago generally faxed resumes to prospective employers. Today, emailed and online applications are much more common. One new program to facilitate mass emailing of correspondence is an automation program by Chehras (2016). Her code will match correspondence to openings based on location and date, generate an email, attach the correspondence, and send the email including delays as desired. Crabtree's Chapter 9 in this volume also discusses email audit studies.

6.2 Matching Responses to Correspondence

Once the experiment has been planned, the participants chosen, the correspondence generated, and the correspondence sent to the participants, the experimenter will still need to match the participants' responses to the characteristics of the correspondence. In a laboratory experiment, this matching can be automated because the experimenter can directly collect the responses from the participant. However, when doing a field experiment, the response can be at some remove from the stimulus. Virtual voice-boxes, PO boxes, and email addresses are common ways of collecting responses and should be chosen with external validity concerns in mind.[5] With generous resources or with a limited number of templates, each stimulus would have its own unique phone number and email address and thus the responses would be directly connected to the correspondence. With more limited resources, it is possible to bin responses based on the main variable combinations of interest, for example, a researcher looking at the effects of race for different 10 year age intervals by gender could have a separate phone number or email address for each age interval*race*gender combination. A drawback of binning rather than doing exact matching is that because correspondence is not directly matched to its response it is difficult to explore the effect of any variables that were not used to create the bins. Without making separate bins by characteristics, it is necessary to match the resumes to the responses using clues from the responses. However, this is costly in terms of person-hours and is not always possible when, for example, firms call back from a number unrelated to the one in the advertisement and do not provide any other identification. Even with binning, it may be difficult to determine when the same company is calling back multiple times in response to the same application.

6.3 Data storage

If possible, keep a copy of everything pertaining to the experiment. In these days of inexpensive storage, it is better to have unused data than to need something and realize it was not preserved and is no longer available. As an example of data size, three thousand resumes, including all the data plus images of the resumes, can take under three hundred megabytes. Each resume's pertinent features must be saved for use in the analysis, commonly stored as variables and a codebook mapping those variables to the resume text. Saving a copy of exactly what is sent to the participant is also a good idea to be able to answer any questions that may arise about what the participants actually received.

Additionally, save the template or process used to generate the submission material. For the Resume Randomizer program, these files consist of scripting commands that detail which inputs should be chosen with specific probabilities and matching constraints. By saving this information, if there are any questions about how the resumes were supposed to be generated, those can be quickly answered. As an example, after the study is run there could be a question about what probabilities were intended during resumes were created for the years of high school graduation. While the variables and codebook can detail what resumes were actually generated, the template is necessary to know the process that generated them. Furthermore, the template can be used as a starting point for future experiments.

[5] Note that researchers using their own domain, such as those from hostgator, can quickly create hundreds of email addresses all with the same passwords and settings. Additionally, Neumark et al. (2016) populated voicemail bins such that each voicemail only had one version of each first name and last name used, which helped with matching. "So if a bin got a call, and they said, 'Hi Jennifer, we'd like to interview you,' then we knew the exact applicant since there was only one Jennifer in that bin," (personal communication, Patrick Button, October 20, 2016).

The final recommendation regarding data storage is to store an off-site backup of everything in case of hard drive failure, fire, or natural disaster. For those who do not have secure online back-ups available from their place of work, Amazon currently sells unlimited storage via Amazon Drive for sixty dollars per year, and a variety of other companies offer similar storage services (e.g., Google Drive, DropBox, iCloud, OneDrive). Sharing data with other researchers after publication at a site such as ICPSR will also protect from data loss. In doing so, be mindful of appropriate data-protection/anonymization protocols and any restrictions imposed by IRB or any governing body for the data (see Gaddis Chapter 3 of this volume for more discussion of IRB concerns).

6.4 If you need to change the resumes mid-experiment

Sometimes correspondence will need to be changed mid-experiment. For example, summary statistics or initial analysis can indicate that a mistake was made in the template(s). Inaccurate calculations of numbers/ages/years, using an outdated version of the template, or completely omitting a section of the resume are all examples of unintentional actions/inactions that might substantially reduce external validity. Alternatively, even after a careful pilot study, unexpected events or findings after the experiment has started can encourage researchers to make modifications to the study. This chapter encourages (and facilitates) mindful preparation, but unforeseen and unavoidable occurrences happen and can lead to the decision to make a correspondence revision mid-experiment despite the reduction in power that comes from dividing the samples.

Mid-experiment revision leads to data storage and data connection challenges. The first challenge is keeping track of the data from resume inputs. For simple designs that use a limited number of templates matched by hand or via mail merge, it is sufficient to mark the resumes before and after the change. Researchers using our program (as of this writing) to create correspondence will end up with two separate datasets, one from before the change and one from after the change. Depending on the change that has been made, the variable names or values may no longer map to each other. Researchers should then post process these two datasets separately before combining in order to match the correct variables together. The second challenge is that responses to the new correspondence need to be identifiable from responses to the old correspondence. If using bins for response collection, that separation may require new email addresses or phone numbers. For more complicated matching procedures it may be sufficient just to keep track of the date at which the change was made. Finally, it is important to keep a clear record of any changes made and when they were made. On hard drives, it is helpful to keep the new data (template, codebook, variable files, etc.) in a separate folder from the pre-revision data to avoid any confusion or lost data. Obviously, a researcher should also clear the changes in correspondence with their IRB if required to do so.

7 Analysis Concerns

The choice of dependent variable will vary by study. In resume audits, the choice between call-back (when the company sends any non-negative response) versus interview (when the company specifically requests an interview) is a common one. There does not seem to be a consensus on which numbers to present, and in our opinion researchers should present both for comparability

across studies. Researchers using other types of correspondence audits should use what is most common in their specific literature unless there is a strong theoretical reason not to.

When multiple stimuli are sent to the same participant (ex. multiple resumes are sent to the same want-ad), it is important to account for between observation (intra-class) correlation. In that case, one should cluster on participant in a regression framework (Lahey and Beasley 2009). For many cases, simply clustering on participant will be sufficient, however some studies may require more complicated methods of correcting standard errors. Clustering can be nested, but if non-nested clusters exist (e.g., different participants sampled over time), traditional cluster inference can only handle one of the dimensions (Cameron and Miller 2015). Alternatively, random effects modeling is commonly used in the metrics of panel data and can be used if the group coefficients are assumed to be uncorrelated with observed group covariates. Both random effects modeling and the more general multilevel modeling (MLM, also called "mixed" models), can handle multiple levels of correlation (e.g., state and participant). A detailed discussion of these different ways of dealing with clustered data is beyond the scope of this chapter, but a good place for interested readers to start is the UCLA Institute for Digital Research and Education webpage on analyzing correlated data (http://www.ats.ucla.edu/stat/stata/library/cpsu.htm).

The related question of when to use participant fixed effects is non-trivial. When sending multiple resumes to a firm, it is tempting to use firm or job opening fixed effects to control for firm characteristics, all items on matched resumes that are matched, and even the point in the business cycle at which the resumes were sent. However, using participant fixed effects when the dependent variable is binary and the researcher is using logit or probit analysis leads to a mis-estimation of the level of differential treatment because it drops all instances where the stimuli were treated the same, leading to the standard Heckman critique (Heckman 1998).

A second Heckman critique about audit discrimination studies is that the magnitude of market discrimination that these studies find has no real meaning because the treatment and control are equivalent by design (Heckman 1998). Thus discrimination magnitudes can only be compared across audit studies but have no real world relevance other than their sign and significance. Neumark (2012) provides a clever method of translating the results from a discrimination audit study into meaningful numbers while Lanning (2013) proposes a method to translate audit-pair findings into wage differentials.

Existence and magnitude of discrimination are not the only outcome of interest even in a discrimination correspondence study. A primary benefit of the larger sizes and better technology with modern correspondence studies is that they are no longer limited to addressing the question, "Is there differential treatment?" and now can start to answer questions of, "Why is there differential treatment?" and "Which sub-groups are most affected?" Pedulla in Chapter 11 of this volume goes more into detail about these important theoretical questions. A simple interaction with main effects can be used to test both of these types of questions. One caveat is that interactive effects require larger sample sizes to find significance at a reasonable power, and researchers should be cognizant of these requirements.

One specific avenue of interest may be testing differential effects by the "quality" of the correspondence. Rather than having the researcher decide what items constitute high quality vs. low quality, it is best to let the market decide what items they prefer. A simple way to get predicted quality is to regress the outcome measure on all items that vary absent the ones that you care about, for example, regress call-back outcomes on all resume items except name (which indicates race/gender), or on all resume items except high school graduation date (which indicates age). Then the predicted Y would be the quality measure absent the variable of interest.

8 Beyond standard Audit studies

Although we have motivated much of this chapter with resume audits, the correspondence review technology does not need to be limited to employment audits. This technology can be expanded to many types of laboratory or natural field experiments (Harrison and List 2004). For example, there is no reason hypothetical correspondence cannot be used with subject pools like Amazon's Mechanical Turk (see Chapter 10 in this volume by Kugelmass for more discussion of Mechanical Turk) or used in conjunction with a natural experiment as in Agan and Starr (2016). The technology can be combined in a laboratory setting with surveys, eye-tracking (ex. Lahey and Oxley 2016), IAT tests (ex. Rooth 2010), other types of laboratory experiments, and so on, to get a richer understanding of what motivates people's choices. Much of this technology has historically been used to explore discrimination in markets, but it does not need to be limited to employment, mortgage markets, or purchasing (Bertrand and Duflo 2016, Neumark 2016). Potential future avenues could include experiments looking at soliciting donations, responses to consumer complaints or political concerns, or the effects of advertising. The use of these methods are only limited by ethical concerns and the researcher's imagination.

9 Technical checklist

- ☐ Determine an externally valid (unobtrusive) signal for the treatment(s)
- ☐ Talk with practitioners and explore current practices in your market
- ☐ Decide on a participant pool
- ☐ Choose how to gather representative inputs
- ☐ Plan response collection method (e.g., email addresses)
- ☐ Review design choices with respect to the expected external validity
- ☐ Get IRB approval for pilot and run a pilot study (optional)
- ☐ Estimate necessary sample size from pilot, previous research, or default estimates
- ☐ Decide on length of time to field experiment
- ☐ Decide on data storage including off-site back-ups and regular back-up schedule while experiment is running
- ☐ Register experiment (optional)
- ☐ Get IRB approval
- ☐ Generate correspondence
- ☐ Submit correspondence
- ☐ Collect participant responses
- ☐ Match responses to correspondence
- ☐ Mid-experiment analysis, revision, and IRB changes (optional)
- ☐ Do primary data analysis as specified in registration, grant proposal, or other initial plan
- ☐ Do exploratory secondary data analysis

References

Agan, AY, Starr, SB (2016) Ban the Box, Criminal Records, and Statistical Discrimination: A Field Experiment. U of Michigan Law & Econ Research Paper No. 16-012. Available at SSRN: https://ssrn.com/abstract=2795795

Baert S, Dieter V (2014) Unemployment or Overeducation: Which is a Worse Signal to Employers?

Barron JM, Bishop J, Dunkelberg WC (1985) Employer search: the interviewing and hiring of new employees. The Review of Economics and Statistics 67(1): 43-52

Bendick Jr M, Brown LE, Wall K (1999) No foot in the door: An experimental study of employment discrimination against older workers. Journal of Aging & Social Policy 10(4): 5-23

Bertrand M, Duflo E, (2016) Field experiments on discrimination via National Bureau of Economic Research. http://www.nber.org/papers/w22014 Accessed 13 Oct 2016

Bertrand M, Mullainathan S (2004) Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. The American Economic Review 94(4): 991-1013

Cameron AC, Miller DL (2015) A practitioner's guide to cluster-robust inference. Journal of Human Resources 50(2): 317-372

Carbonaro W, Schwarz J (this volume) Which jobs, which employers, which labor markets?: Challenges in designing and conducting a labor market resume study, In: Gaddis S An Introduction to Audit Studies, 1st edn. Springer

Charness G, Gneezy U, Kuhn MA, (2012) Experimental methods: Between-subject and within-subject design. Journal of Economic Behavior & Organization 81(1): 1-8

Chehras N (2016) Automating correspondence study applications with python and SQL: Guide and code. *Mimeo*

Cohen J (1977) Statistical power analysis for the behavioral sciences. Academic Press. New York

Cohen, J (1992) A power primer. Psychological Bulletin. 112(1): 155-159

Crabtree C (this volume) Social Science by Email: Adapting and Improving Email Audit Studies, In: Gaddis S An Introduction to Audit Studies, 1st edn. Springer

Darolia, R, Koedel C, Martorell P et al. (2015) Do employers prefer workers who attend for-profit colleges? Evidence from a field experiment. Journal of Policy Analysis and Management 34(4): 891-903

Deming DJ, Yuchtman N, Abulafi A et al. (2016) The value of postsecondary credentials in the labor market: An experimental study. The American Economic Review 106(3): 778-806

Friedman S, Reynolds A, Scovill S (2013) An Estimate of Housing Discrimination Against Same-Sex Couples. Available via the US Department of Housing and Urban Development: http://big.assets.huffingtonpost.com/hud.pdf

Gaddis S (2014) Discrimination in the credential society: An audit study of race and college selectivity in the labor market. Social Forces 95(1): 1-29

Gaddis S (this volume) The ethics of audit studies: best practices and academic cooperation, In: Gaddis S An Introduction to Audit Studies, 1st edn. Springer

Gaddis S (this volume) By any other name? Signaling characteristics in correspondence audits, In: Gaddis S An Introduction to Audit Studies, 1st edn. Springer

Gulliford MC, Obioha UC, Chinn S (1999) Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: data from the Health Survey for England 1994. American Journal of Epidemiology 149(9): 876-883

Harrison GW, List JA (2004) Field experiments. Journal of Economic literature 42(4): 1009-1055

Heckman JJ (1998) Detecting discrimination. The Journal of Economic Perspectives 12(2): 101-116

Holzer HJ (1996) What employers want: Job prospects for less-educated workers. Russell Sage Foundation, New York

Horton JJ (forthcoming) The effects of algorithmic labor market recommendations: Evidence from a field experiment. Journal of Labor Economics

Howden D (2016) Interviews per hire: recruiting KPIs. Available via Workable. https://resources.workable.com/blog/interviews-per-hire-recruiting-metrics/ Accessed 13 Oct 2016

Jobs Opening and Labor Turnover Survey (JOLTS). Bureau of Labor Statistics. http://www.bls.gov/jlt/home.htm. Accessed 13 Oct 2016

Kang SK, Decelles KA, Tilcsik A et al. (2016) Whitened résumés: race and self-presentation in the labor market. Administrative Science Quarterly 61(3): 469-502

Kugelmass H (this volume) Technology and Audit Studies: A Return to Phone Audits and Using Mechanical Turk for Pre-Tests, In: Gaddis S An Introduction to Audit Studies, 1st edn. Springer

Lahey JN (2008) Age, women, and hiring an experimental study. Journal of Human Resources 43(1): 30-56

Lahey JN, Beasley RA (2009) Computerizing audit studies. Journal of Economic Behavior & Organization 70(3): 508-514

Lahey JN, Oxley D (2016) Discrimination at the intersection of age, race, and gender: Evidence from a lab-in-the-field experiment. Working Paper

Lanning JA (2013) Opportunities denied, wages diminished: Using search theory to translate audit-pair study findings into wage differentials. BE Journal of Economic Analysis and Policy 13(2): 921-958

Maas CJ, Hox, JJ (2005) Sufficient sample sizes for multilevel modeling. Methodology 1(3): 86-92

Maurer R (2016) More employers moving to fewer interviews. Available via Society for Human Resources Management. Accessed 13 Oct 2016

Moynihan LM, Roehling MV, LePine MA et al (2003) A longitudinal study of the relationships among job search self-efficacy, job interviews, and employment outcomes. Journal of Business and Psychology 18(2): 207-233

Neumark D (2012) Detecting discrimination in audit and correspondence studies. Journal of Human Resources 47(4): 1128-1157

Neumark, D (2016) Experimental research on labor market discrimination. NBER working paper series. http://www.nber.org/papers/w22022 Accessed 17 Oct 2016

Olian JD, Schwab DP, Haberfeld Y (1988) The impact of applicant gender compared to qualifications on hiring recommendations: A meta-analysis of experimental studies. Organizational Behavior and Human Decision Processes 41(2): 180-195

Olken BA (2015) Promises and perils of pre-analysis plans. The Journal of Economic Perspectives 29(3): 61-80

Oreopoulos P (2011) Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. American Economic Journal: Economic Policy 3(4): 148-71

Pager D, Pedulla DS (2015) Race, self-selection, and the job search process. American Journal of Sociology, 120(4): 1005-1054

Pedulla D (this volume) Emerging frontiers in audit study research: Mechanisms, variation, and generalizability, In: Gaddis S An Introduction to Audit Studies, 1st edn. Springer

Phillips C (2016) Do comparisons of fictional applicants measure discrimination when search externalities are present? Evidence from existing experiments. Working Paper

Rooth DO (2010) Automatic associations and discrimination in hiring: Real world evidence. Labour Economics, 17(3):523-534

Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. Science 211 (4481): 453-458

Vuolo M (this volume) To match or not to match? Statistical and substantive considerations in audit design and analysis, In: Gaddis S An Introduction to Audit Studies, 1st edn. Springer