Do Comparisons of Fictional Applicants Measure Discrimination When Search Externalities Are Present? Evidence from Existing Experiments

David C. Phillips[1]
Wilson Sheehan Lab for Economic Opportunities
Department of Economics
University of Notre Dame

February 2017

Abstract
Researchers commonly use fictional applicants to measure discrimination. However, such experiments can confound discrimination against an individual's characteristics with employers' responses to the composition of the applicant pool. Such confounding occurs when one applicant's characteristics affect another applicant's likelihood of success. I find evidence of such spillovers using data from several existing experiments. Applicants randomly assigned to compete against higher quality applicant pools receive more callbacks. Under one reasonable set of assumptions, adjusting for applicant pool composition increases measured discrimination by 19% on average. Such confounding can be eliminated by avoiding experimental designs that stratify treatment assignment by vacancy.

JEL Codes: J71; J64; C93

Keywords: discrimination; correspondence experiments; experimental design

[1] E-mail: David.Phillips184@nd.edu. I have benefitted from comments from Peter Brummund, Patrick Button, Sarah Estelle, Joanna Lahey, Steve McMullen, Philip Oreopoulos, Dan-Olof Rooth, Russell Toth, and Daryl Van Tongeren as well as seminar participants at Georgetown University and the University of Alabama. David Green, Brennan Mange, and Molly Meyer provided excellent research assistance. This study received financial support from Hope College, my previous employer. I am grateful to the many authors and journals that have made data publicly available. All remaining errors and omissions remain the responsibility of the author.

1. Introduction

Researchers frequently measure discrimination with fictional applications to real vacancies in employment and rental housing. Such experiments control and randomly assign all observed characteristics of a résumé. Applicants treated with a characteristic of interest, e.g. a black-sounding name, appear otherwise similar to applicants without the characteristic. Measuring discrimination reduces to a simple comparison of callback rates: which group of fictional applicants is more likely to be called in for a job interview or apartment showing? The simplicity and rigor of such correspondence experiments make these studies perhaps the most convincing, common method for measuring discrimination. Experimenters have exploited this research design to measure discrimination by ethnicity (Bertrand and Mullainathan, 2004; Oreopoulos, 2011; Booth, et. al., 2011; Hanson and Hawley, 2011; Arceo-Gomez and Campos-Vazquez, 2014; Ewens et. al., 2014; Baert, et. al., 2015), age (Lahey, 2008, Neumark, et. al., 2015), residential location (Phillips, 2015), ongoing unemployment duration (Kroft, et. al., 2013; Eriksson and Rooth, 2014), post-secondary credential (Deming, et. al. 2016), and many other dimensions. Existing critiques (e.g. Heckman, 1998) have identified potential flaws in interpreting the results of correspondence experiments as clear evidence of discrimination, though the literature has generally responded with methods robust to such well-documented critiques (e.g. Neumark, 2012; Neumark and Rich, 2016). In the present study, I develop a new critique. I show that differences in callback rates from correspondence experiments can confound the direct effect of discrimination with employer responses to the composition of the applicant pool. I test empirically for this phenomenon in several existing studies and find it to be significant, affecting the interpretation of existing results and warranting changes to optimal experimental design in future studies.

Experiments confound discrimination with applicant pool composition when simple and common experimental design choices interact with spillovers across applicants. In early studies,

researchers created "matched pairs" of real applicants who were similar on all characteristics except the treatment. Modern experiments use fictional people, large samples, and electronic correspondence but often maintain a similar design. They send 2 or more applications to each vacancy with treatment and control split evenly within the vacancy, i.e. stratifying treatment by vacancy. However, such designs mechanically correlate treatment of a given applicant with the characteristics of the applicant pool. In an experiment sending 2 white names and 2 black names to each job, a candidate with a black name will compete against 2 white names while a candidate with a white name will compete against only 1 white name. If an employer's perception of the quality of the other applicants matters for one's own callback rate, then simple differences in callback rates will include applicant pool composition effects. Formally, I show that simple differences in callback rates will not identify the direct effect of discrimination when three conditions all hold: researchers send multiple applicants to each vacancy, researchers stratify treatment (e.g. use "matched pairs"), and an employer's evaluation of one applicant depends on the characteristics of other applicants. Many experiments meet the first two characteristics; hence, the interpretation of existing experiments hinges on the third condition, whether spillovers exist in practice.

Simple theory can motivate both positive and negative spillovers across applicants. Negative spillovers may occur if one good applicant crowds a second candidate out of a congested interview schedule. Positive spillovers may arise if employers infer the quality of one applicant from another applicant's characteristics. For instance, a prejudiced employer may wish to avoid interacting with black applicants but cannot observe race perfectly before the interview. The presence of one application with a black-sounding name may cause concern about the racial composition of the entire pool, leading the employer to re-post the ad later. I develop a simple, formal model including both negative spillovers from displacement and positive information spillovers. Theory rationalizes both negative and positive spillovers across applicants.

As my main contribution, I establish the empirical importance of spillovers across applicants. Estimating the effect of applicant pool quality on the success of any one applicant provides an empirical challenge. The well-known reflection problem biases simple estimates of peer effects (Manski, 1993). Also, an experiment with sufficient statistical power to detect main treatment effects will not be powered to detect spillovers. Spillover effects should be smaller than direct effects, and average applicant pool quality varies less than individual quality. Both of these features reduce statistical power. Thus, I pool data across several existing correspondence studies and exploit random variation in applicant pool quality generated by these experiments to estimate unbiased, precise spillover effects. I typically cannot test for spillovers using the original treatment variables, e.g. race, because stratification of treatment mechanically and perfectly correlates treatment of one applicant with non-treatment of another applicant. However, some recent experiments assign other characteristics on the application randomly and independently of the characteristics of other applicants to the same job. These ancillary applicant characteristics provide random variation in the quality of the applicant pool that is unrelated to the characteristics of a given applicant. Because I exploit experimental random assignment, I can estimate the causal effect of one applicant's characteristics on another applicant's success rather than the effect of fixed vacancy characteristics, such as employers that choose to interview many applicants. I use this variation and detect large, positive spillovers across applicants. When an experiment randomly assigns one applicant to have higher quality characteristics, the response rate of other applicants to the same job increases. Averaging across 7 existing experiments, improving one applicant's quality increases the other applicants' response rates by 31% of the benefit to the original applicant.

Large positive spillovers affect the performance of different experimental designs. Common measures of discrimination will be biased[2] in the presence of spillovers. Calculating bias directly through a new study comparing experimental designs would require a prohibitively large sample. Bias can be measured in existing data with stronger parametric assumptions. If spillovers scale linearly in proportion to the direct importance of an application characteristic, stratified designs underestimate racial discrimination by 19%, on average. Stratified experimental designs do provide some benefits by creating better balance and hence greater statistical efficiency, but Monte Carlo simulations indicate these benefits are small relative to the bias. These bias estimates depend on parametric assumptions which are difficult to test. Banerjee, et. al. (2016) argue that randomized experiments prove valuable precisely because they provide evidence robust to audiences with skeptical prior beliefs. This lens can help interpret the bias estimates: I identify reasonable beliefs under which a skeptical observer can expect large bias given the observed magnitude of spillovers. An experimenter facing a skeptical audience would thus avoid stratified and matched pair designs, which can generate large bias under reasonable assumptions.

I provide guidance on how to improve experimental design in the presence of spillovers. Most simply, an experimenter can send one application to each job (e.g. Ewens, et. al., 2014) or multiple applications with treatment assigned independently (e.g. Kroft, et. al. 2013). For a researcher who has already conducted an experiment using stratification, I demonstrate how to bound bias by adding a sub-sample with a non-stratified experimental design. Such bounding can be informative with large samples. Finally, I consider the possibility of correcting measured discrimination using bias approximations as above. However, I find suggestive evidence that the

[2] Throughout the paper, I take the point of view of an experimenter who wishes to estimate the direct effect of some applicant characteristic holding the applicant pool constant; hence, I interpret any applicant pool composition effect as bias. This perspective matches the interpretive language employed in most correspondence studies.

parametric assumptions required to estimate bias do not always hold. The actual direction and magnitude of bias may be difficult to predict in practice.

In the remainder of the paper, I further develop this argument. Section 2 describes the literature on correspondence studies of discrimination, provides a simple theory to motivate spillovers in the context of correspondence studies, and formally defines the empirical problem that spillovers pose for measures of discrimination. Section 3 describes my empirical strategy, the data, and how that strategy maps to the data. Section 4 presents results on the extent and magnitude of spillovers across applicants in existing studies. Section 5 estimates the extent of bias caused by spillovers and the tradeoff with efficiency. Section 6 provides guidance on designing experiments robust to spillovers. Section 7 concludes.

2. Correspondence Experiments in the Context of Spillovers

2.1. Background on Correspondence Experiments

Discrimination proves difficult to measure because so many unobservable or difficult to measure characteristics correlate with race. Audit experiments were created to fill this void. Earlier studies (Yinger, 1995; Riach and Rich, 2002) as well as recent additions (e.g. Pager, et. al. 2009) recruit pairs of actual people, give them fictional identities, and have them apply to actual jobs or apartments. Researchers match pairs to be similar on various characteristics except one dimension, e.g. one white and one black tester who otherwise appear similar and are given similar fictional credentials. One can then compare differences in positive outcomes, such as being called back for an interview, to measure discrimination. By making pairs similar, matched-pair audits help eliminate many concerns about unobservable variables.

However, audit studies may fail to measure discrimination if testers differ on some unmatchable dimension (Heckman, 1998). Correspondence studies (e.g. Bertrand and Mullainathan, 2004) solve this concern by using fictional applicants applying to vacancies

electronically or by mail. Applying from a distance allows experimenters to control all information seen by the vacancy, and large samples combined with randomization of fictional applicant characteristics allow experimenters to balance all observed characteristics between treatment and control groups. Given its attractive features, a large number of recent studies use this method, measuring discrimination as the difference in the probability of receiving a positive communication for applicants with different apparent ethnicity (Bertrand and Mullainathan, 2004; Oreopoulos, 2011; Booth, et. al., 2011; Hanson and Hawley, 2011; Arceo-Gomez and Campos-Vazquez, 2014; Ewens et. al., 2014), age (Lahey, 2008; Neumark, et. al. 2015), residential location (Phillips, 2015), ongoing unemployment duration (Kroft, et. al., 2013; Eriksson and Rooth, 2014), and more.

While correspondence studies allow experimenters to better match applicants, some critiques remain. These studies cannot in general distinguish taste-based discrimination from statistical discrimination based on either the mean or variance of unobservable characteristics (Heckman, 1998). This critique, though, is widely known and some statistical tests (Neumark, 2012) and clever experimental designs (Ewens, et. al., 2014) have been developed to address this concern. Correspondence experiments have become a prominent empirical tool for detecting discrimination, but few, if any, studies have considered whether spillovers across applicants affect the correct interpretation of results from such experiments.

2.2. An Illustrative Theory of Spillovers in Correspondence Experiments

A very large literature discusses the theory (Pissaridies, 2000; Hosios, 1990) and empirics (Blundell, et. al. 2004; Dahlberg and Forslund, 2005; Ferracci, et. al. 2014; Pallais, 2014; Gautier, et. al. 2012; Lalive, et. al. 2015; Albrecht, et. al. 2009; Crépon, et. al. 2013) of search externalities and peer effects (Manski, 1993). I build a simple theoretical model to frame my empirical analysis, describing reasons for spillovers across applicants in the context of correspondence studies. I will focus on the extent to which the characteristics of one applicant to a job can affect the probability

that another applicant to the same employer receives an interview request. For clarity, I define and interpret a specific model with few formal details. The appendix provides a formal exposition, and the end of this section considers how other spillover theories can fit into the stated model.

Suppose that an employer wishes to interview applicants for vacancy j. With increasing costs of interviewing the marginal applicant, interview space becomes scarce and applicants will compete based on their characteristics. For concreteness, suppose that marginal costs rise very quickly. The employer schedules at most one interview.[3] Any given applicant receives an interview only if he[4] is the best applicant *and* if the employer views the value of interviewing him as greater than the cost of doing so. Formally, let Y_{ij} indicate whether applicant i receives an interview from job j. The probability of receiving an interview is:

$$\Pr\left[Y_{ij} = 1 \middle| \boldsymbol{X_j}\right] = \Gamma\left(\boldsymbol{X_j}\right) * \Psi\left(\boldsymbol{X_j}\right) \qquad (1)$$

X_{ij} is a characteristic of interest (e.g. racial connotation of the name), $\boldsymbol{X_j}$ is the vector of this characteristic for all applicants to job j, $\Gamma\left(\boldsymbol{X_j}\right)$ is the probability that person i is the best applicant in the pool, and $\Psi(\boldsymbol{X_j})$ is the probability that the value of applicant i exceeds the cost of the interview, i.e. that the employer interviews anyone. Consider a characteristic X_{ij} that employers view positively. An applicant assigned a high value of X_{ij} will receive more callbacks both because he is more likely to be the best applicant, $\frac{\partial \Gamma(\mathbf{X_j})}{\partial X_{ij}} > 0$, and because his good characteristic makes the interview more likely to be worth the cost for the employer, $\frac{\partial \Psi(\mathbf{X_j})}{\partial X_{ij}} > 0$.

However, changing the characteristics on one person′s application may also affect other applicants to the same job. There are two main spillovers onto person i if we improve the

[3] Any number n less than the total number of applicants gives similar comparative statics.
[4] For clarity, I will refer to the employer with female pronouns and the applicant with male pronouns.

characteristic X_{kj} for some other person $k \neq i$. First, there is a displacement effect:

$$\frac{\partial \Gamma(\mathbf{X_j})}{\partial X_{kj}} < 0, \ \forall \ \ k \neq i \qquad (2)$$

Improving a competing applicant's characteristic X_{kj} potentially displaces applicant i out of a fixed number of interview slots. While intuitive, crowd out may be less likely in correspondence experiments if employers refill the interview spots vacated by high quality, fictional applicants who disappear. Second, a higher quality pool may increase the employer's likelihood of interviewing at all:

$$\frac{\partial \Psi(\mathbf{X_j})}{\partial X_{kj}} > 0, \ \forall \ \ k \neq i \qquad (3)$$

More generally, a higher quality pool may induce an employer to schedule more interviews for marginal applicants. See the appendix for a more formal treatment.

Consider a concrete example. Suppose an employer views residence in one particular neighborhood as a signal of unreliability. When the employer receives applications from that neighborhood in response to a particular job ad, the employer conducts classic statistical discrimination (Phelps, 1972) and discounts such applicants. The presence of these other applicants from 'bad' neighborhoods may also spillover onto our applicant of interest. First, he faces weaker opponents, which improves the odds of ranking high enough to obtain an interview slot. On the other hand, the employer may view the presence of applicants from 'bad' neighborhoods as a negative signal about the entire applicant pool, including our applicant of interest. Perhaps a temp agency near the 'bad' neighborhood has directed many low quality clients to apply, only some of whom can be identified by their addresses. Given these concerns, the employer discards the current

pool and re-posts the ad at a different time or on a different job board.[5] In summary, improving a positive characteristic for one applicant may positively or negatively affect other applicants.

In the empirical section I find evidence of positive spillovers. The exposition above and in the appendix provides one simple rationalization of the results: an employer with limited information about any given applicant uses the quality of other applicants in the pool to generate a more accurate evaluation. However, the above model provides predictions identical to various other mechanisms of positive spillovers. Different specifications of how an employer evaluates an individual's productivity based on group information could represent either rational statistical learning (e.g. Banerjee, 1992) or behavioral responses to the same data, such as priming and halo effects. If one reinterprets the option to not hire an applicant as a decision to hire from another source or at a different time, the model becomes a theory of a rational employer who searches across sources or dynamically. If one allows not just the quality but also the quantity of the applicant pool to vary, then positive spillovers can result from the "market thickness" effect found in canonical labor market search and matching models (Pissarides, 2000). Many simple and standard theories predict positive spillovers in correspondence studies.

2.3. What Do Differences in Callback Rates Measure in the Presence of Spillovers?

The interpretation of differences in callback rates changes in the presence of spillovers. Studies typically measures a difference in response rates with a simple dummy variable regression:

$$Y_{ij} = \alpha + \theta T_{ij} + v_{ij} \quad (4)$$

where Y_{ij} measures a positive response by employer j to the application of person i (e.g. a dummy for whether the applicant receives an interview offer or other positive response), T_{ij} is a treatment

[5] Similarly, a prejudiced employer may wish to avoid interacting with black applicants but cannot observe race perfectly before interviews. The presence of a stereotypically black name on one application may lead the employer to re-post the ad elsewhere.

dummy (e.g. a black-sounding name), and v_{ij} is an error term. In this context, OLS measures the difference in callback rates between treated and control, $\hat{\theta} = \bar{Y}^T - \bar{Y}^C$. Given that the experiment balances other observable variables across treatment and control, the literature typically interprets the size and significance of $\hat{\theta}$ as evidence of discrimination.[6]

However, in the presence of spillovers across applicants, differences in callback rates may combine the direct effect of treatment on the treated applicant with the effect of changing the composition of the applicant pool. For equation (4) to measure only the direct effect of treatment, the researcher must make the standard stable unit treatment value assumption (Cox, 1958) that the effects of treatment do not spill over onto the controls. However, for practical reasons most experiments send multiple fictional job applications to the same vacancy, and as shown in the theory section above, the characteristics of one applicant may affect another applicant's success rate. Suppose that while (4) represents the model estimated by the researcher, a model with spillovers represents the true model:

$$Y_{ij} = \alpha + \beta T_{ij} + \delta\beta(K-1)\bar{T}_{(i)j} + \epsilon_{ij} \quad (5)$$

Callbacks respond to both the applicant's own characteristics and the mean of other applicants' characteristics. For a racial name treatment, the term $\bar{T}_{(i)j}$ counts the proportion of other applicants to vacancy j other than person i who list a black-sounding name. In the appendix, I show formally that a simple statistical discrimination model with correlated unobservable characteristics across applicants implies equation (5). I will refer to β as the "direct effect" of treatment and δ as the parameter determining the relative direction of "spillover effects" or "applicant pool composition

[6] Such studies do not generally distinguish between taste-based and statistical discrimination, and this fact complicates interpretation of audit studies. But this point has already been well-documented in the literature (Heckman, 1998; Neumark, 2012).

effects." K is the number of applicants per job. In the language of Manski (1993), $\delta\beta(K-1)$ measures "exogenous" peer effects.

In an experiment that *randomly and independently* assigns treatment status, a simple difference in callback rates measures only the direct effect of treatment. Suppose that the researchers find a sample of vacancies and randomly assign one applicant to each vacancy. In this context the difference in callback rates measures:

$$E\left[Y_{ij}\middle|T_i=1,\bar{T}_{(i)j}=\bar{T}\right]-E\left[Y_{ij}\middle|T_i=0,\bar{T}_{(i)j}=\bar{T}\right]=\beta$$

where $\bar{T}$ is the average proportion of treated applicants other than person i in the population. Thus, differences in callback rates estimate the direct effect of treatment when the experimental design sends only one applicant to each vacancy. This fact remains true if the experiment assigns multiple applications to each vacancy but selects the treatment status of each application independently.

However, in common research designs differences in callback rates combine the direct effect of treatment with an applicant pool composition effect. Many experiments send *"matched pairs"* or more generally randomize treatment status but *stratify by vacancy*. For instance, Bertrand and Mullainathan (2004) send 4 applications to each job. They randomly assign which applications receive black or white names, but each job receives exactly two black and two white names. As a result, simple differences in callback rates measure the effect of switching an applicant from white to black while also switching another member of the applicant pool from black to white. For a vacancy receiving K applications, the difference in callback rates measures:

$$\begin{gathered}E\left[Y_{ij}\middle|T_i=1,\bar{T}_{(i)j}=\bar{T}\right]-E\left[Y_{ij}\middle|T_i=0,\bar{T}_{(i)j}=\bar{T}+\frac{1}{K-1}\right]=\\ \left(E\left[Y_{ij}\middle|T_i=1,\bar{T}_{(i)j}=\bar{T}\right]-E\left[Y_{ij}\middle|T_i=0,\bar{T}_{(i)j}=\bar{T}\right]\right)\\ +\left(E\left[Y_{ij}\middle|T_i=0,\bar{T}_{(i)j}=\bar{T}\right]-E\left[Y_{ij}\middle|T_i=0,\bar{T}_{(i)j}=\bar{T}+\frac{1}{K-1}\right]\right)\\ =\beta-\delta\beta\end{gathered}$$

When experiments use a stratified design to compare callback rates, they combine the direct effect of treatment, β, with the effect of changing the applicant pool, $\delta\beta$.

Such confounding occurs under three conditions. First, the experiment sends multiple applications to each vacancy ($K > 1$). Second, the experiment uses stratified randomization of treatment by vacancy. Third, spillovers across applicants to the same vacancy exist ($\delta \neq 0$). Most correspondence studies meet the first two conditions. However, researchers frequently attribute differences in callback rates to the direct effect of treatment rather than applicant pool composition effects. This interpretation implicitly assumes that no spillovers exist, $\delta = 0$. If spillovers do exist, this will lead to a misinterpretation. Suppose that employers view treatment negatively ($\beta < 0$). If giving a negative attribute to a competitor improves one's own chances by making room on the interview list, $\delta < 0$, then experiments will overestimate the direct effect of treatment. On the other hand, if a negative change to another applicant sours the employer on all applicants from the interview source used by the experiment, $\delta > 0$, then experiments will underestimate the direct effect of treatment. Thus, the importance of the present critique and the interpretation of correspondence studies hinges on an empirical question, whether the characteristics of one applicant affect responses to other applicants.

3. Empirical Strategy

3.1. Reduced Form Model for Measuring Spillovers

I test for the existence and extent of spillovers across applicants using random variation in the composition of applicant pools created by correspondence experiments. I cannot directly estimate equation (5) because stratified randomization assigns treatment such that T_{ij} and $\bar{T}_{(i)j}$ are perfectly collinear. However, many experiments assign at least one applicant characteristic, Z_{ij}, randomly and independently, that is without stratifying randomization. This allows me to test for the existence of spillovers with respect to Z_{ij}. For example, Bertrand and Mullainathan (2004) randomly assign addresses to résumés and provide data on log median household income of the listed address's census tract. Because they do not stratify randomization of this variable, the

composition of addresses in the applicant pool randomly varies from job to job. I can test whether this random variation in other applicants' quality affects a particular applicant's callback probability:

$$Y_{ij} = \alpha + \psi Z_{ij} + \delta\psi(K-1)\bar{Z}_{(i)j} + \zeta X_{ij} + \epsilon_{ij} \qquad (6)$$

As noted above, Z_{ij} is some applicant characteristic that experimenters assign randomly and independently. We are interested in the coefficient on $\bar{Z}_{(i)j}$ which measures spillovers of other applicants' characteristics onto applicant i with the direction of spillovers determined by δ. $\bar{Z}_{(i)j}$ is the mean[7] of the characteristic Z_{ij} for other applicants to the same job. I also include Z_{ij} in the regression. Because Z_{ij} is chosen to be uncorrelated with $\bar{Z}_{(i)j}$, including Z_{ij} is not strictly necessary, but ψ measures whether employers view Z_{ij} positively or negatively which helps interpret the spillover coefficient. Finally, I also include a vector of control variables X_{ij} because some experiments randomly assign Z_{ij} conditional on some controls. For instance, Bertrand and Mullainathan (2004) randomly assign addresses conditional on city (Boston or Chicago). In this regression, I focus on the parameter δ. Provided that the experiment assigns Z_{ij} randomly and independently across applicants, a non-zero δ implies that spillovers across applicants exist.

3.2. Instrumental Variables Model for Measuring Spillovers

At times, I will interpret equation (6) as a reduced form test for the existence of spillovers across applicants that corresponds to an instrumental variables specification. Consider the following two-stage IV specification:

$$\text{First stage:} \quad \bar{Y}_{(i)j} = \phi_0 + \phi_1\bar{Z}_{(i)j} + \phi_2 X_{ij} + \phi_3 Z_{ij} + v_{ij} \quad (7)$$

$$\text{Second stage: } Y_{ij} = \alpha + \delta_{IV}\bar{Y}_{(i)j} + \phi_4 Z_{ij} + \phi_5 X_{ij} + \epsilon_{ij} \quad (8)$$

[7] In a previous version of this paper, I use the sum rather than the mean and obtain similar results. The two options give identical results (subject to re-scaling) if the number of experimental applicants does not vary across vacancies.

The term $\bar{Y}_{(i)j}$ in the second stage, which is the callback rate to other applicants to job j, allows one applicant's success in receiving a response to affect another applicant's probability of receiving a response. OLS using equation (8) likely estimates δ_{IV} with bias, either positive bias because some vacancies simply call back more applicants or negative bias because stratified experimental designs create negative correlation between the quality of two applicants to the same vacancy. Thus, I instrument call backs to other applicants, $\bar{Y}_{(i)j}$, with random variation in those applicants' quality, $\bar{Z}_{(i)j}$, as in equation (7). Most literally, this IV model measures the effect of other applicants' success on the present applicant's success. In the language of Manksi (1993), δ_{IV} measures an "endogenous peer effect." The sign of δ_{IV} indicates the direction of spillovers. However, I will interpret spillovers in the IV model as a simple rescaling of the reduced form results rather than as a structural parameter. In the appendix, I demonstrate that the sign of δ_{IV} in the IV specification versus δ in the reduced form are the same, which is a particular case of a well-known general result linking treatment effects measured by reduced form and IV. Likewise, the literature notes that, in general, exogenous and endogenous peer effects cannot be separately identified (Manski, 1993).

The IV model rescales spillover estimates to have intuitive units that are easily comparable across different studies. By the usual two-stage least squares logic, estimating equation (8) by OLS while substituting predicted callback rates, $\bar{\hat{Y}}_{(i)j}$, from equation (7) will give the same estimate for δ_{IV} as 2SLS. In the correspondence experiment literature (e.g. Bertrand and Mullainathan, 2004), predicted call back rates measure applicant quality. Thus, δ_{IV} can be interpreted as the response of one applicant's callback rate to the quality of other applicants rates. Since all studies use a callback outcome, δ_{IV} measures spillovers in common units across studies.

The IV specification also clarifies the conditions under which my instrument of other applicants' quality is valid. The instrument must be relevant such that $\bar{Z}_{(i)j}$ is correlated with

callbacks to other jobs. It must be exogenous, which is satisfied by random assignment of Z_{ij}. It must not be collinear with Z_{ij}, i.e. not assigned by stratified randomization. These conditions will prove useful in interpreting the reduced form.

The data from some studies provides multiple valid instruments that are randomly assigned without stratification. With multiple exogenous options, I wish to choose the strongest/most relevant instrument. Belloni, Chernozkuhov, and Hansen (2014) provide a procedure for systematically selecting strong instruments. They select instruments out of a set of exogenous instruments using an estimator of the first stage relationship that penalizes model complexity (LASSO). I cannot directly apply their method to the first stage in equation (7). A strong first stage in (7) will mechanically select for a characteristic that affects the applicant listing the characteristic and other applicants to the same job in the same direction.[8] However, I can apply a similar procedure designed to identify the characteristics to which employers most strongly respond. Formally, I apply a LASSO estimator to equation (9):

$$Y_{ij} = \alpha + \psi Z_{ij} + \zeta X_{ij} + \epsilon_{ij} \quad (9)$$

From a vector of exogenous potential instruments Z_{ij}, I select the variable that most strongly predicts employers' responses to the applicant listing the characteristic. I can then test for spillovers across applicants with the strongest available instrument.

3.3. Inference across Multiple Studies

Pooling spillover estimates across studies will prove useful. Spillovers will typically be of smaller magnitude than direct effects, and average quality $\bar{Z}_{(i)j}$ will vary much less than individual

[8] Consider a job receiving applications from persons A, B, C, and D who all list a characteristic that employers view positively. If person A is person i, then the first stage relates callbacks to B, C, and D and the characteristic for B, C, and D. The employers view the characteristic positively, then callbacks to B and the characteristic for B are positively related. But the relationship across all three people will be strongest if the characteristic for person B also spills over to callbacks for C and D. In this way, selecting strong instruments based on equation (7) will bias one toward selecting instruments that have direct and spillover effects in the same direction.

quality Z_{ij}. Hence, sample sizes chosen to detect direct effects will have limited power to detect spillovers in any given study. In practice, pooling yields more precise spillover estimates.

A fully interacted model can test for spillovers jointly across studies. Consider modifying equations (6), (7), and (8) to be fully interacted models estimated on a pooled sample:

$$Y_{ij} = \sum_s I^s * \left[\delta^s \psi^s (K^s - 1)\bar{Z}^s_{(i)j} + \psi^s Z^s_{ij} + \zeta^s X^s_{ij} + w^s\right] + \epsilon_{ij} \qquad (6')$$

$$\bar{Y}^s_{(i)j} = \sum_s I^s * [\phi^s_1 \bar{Z}^s_{(i)j} + \phi^s_3 Z^s_{ij} + \phi^s_2 X^s_{ij} + w^s] + v_{ij} \quad (7')$$

$$Y_{ij} = \sum_s I^s * [\delta^s_{IV} \bar{Y}^s_{(i)j} + \phi^s_4 Z^s_{ij} + \phi^s_5 X^s_{ij} + w^s] + \epsilon_{ij} \quad (8')$$

Equations (6'), (7'), and (8') differ from the single-study specification in three ways. First, I combine the samples from all studies and denote variables from a particular study with superscript s, e.g. X^s_{ij} indicates the control variables from study s. Second, I include study dummies I^s with coefficients w^S. Third, I interact the study dummies with all other variables, allowing for study-specific coefficients.[9] These fully-interacted models exactly replicate the study-by-study coefficients but can also test for spillovers jointly across the studies.

With slight modifications, the regressions above can also average spillovers across studies:

$$Y_{ij} = \delta\psi(K-1)\bar{Z}_{(i)j} + \psi Z_{ij} + \sum_s I^s * \left[\zeta^s X^s_{ij} + w^s\right] + \epsilon_{ij} \qquad (6'')$$

$$\bar{Y}_{(i)j} = \phi_1 \bar{Z}_{(i)j} + \phi_3 Z_{ij} + \sum_s I^s * [\,\phi^s_2 X^s_{ij} + w^s] + v_{ij} \quad (7'')$$

$$Y_{ij} = \delta_{IV} \bar{Y}_{(i)j} + \phi_4 Z_{ij} + \sum_s I^s * [\phi^s_5 X^s_{ij} + w^s] + \epsilon_{ij} \quad (8'')$$

These specifications retain study fixed effects and their interactions with all study controls; thus, they continue to identify spillovers within each study. Two changes appear. First, I combine each of the study specific quality instruments Z^s_{ij} into a standardized variable Z_{ij}. I transform the study-

[9] This sets control variables to zero when they come from a different study than the observation of interest. Because I include study fixed effects, choosing a different value from zero leads to identical estimates on all parameters of interest.

specific quality instruments into a "standard deviations above the mean" z-score within each study so that they have comparable magnitudes. For instruments that are negative signals (e.g. smokers applying to housing), I reverse the sign. Second, I restrict the spillovers coefficients to take a common value across studies. As in a typical study of heterogeneous effects, a common coefficient measures a weighted average of the study-specific effects. Averaging over a variety of contexts estimates spillovers with greater precision and external validity than any one study.

3.4. Data

I use data from several correspondence experiments. I focus on collecting high-quality studies for which data is available. Two research assistants completed separate literature searches. These searches included examining articles on Google Scholar citing Bertrand and Mullainathan (2004) and searching for the terms "audit experiment," "résumé experiment," "discrimination experiment," "correspondence experiment," "correspondence study," and "audit study" using the search functions of several top journals' websites[10] and Google. I include studies resulting from these searches only if I can implement my empirical strategy. I require that the study has publicly available data (or data that I collected), sends multiple applications to each vacancy, and has some instrument Z_i which the researchers assign randomly without vacancy stratification. These criteria limit the sample significantly. For example, I exclude many studies without public data, Ewens, et. al. (2014) because they only send one application per vacancy, and Arceo-Gomez and Campos-Vazquez (2014) because all randomly-assigned variables in the available data are stratified.

Appendix Table 1 alphabetically lists the studies from which I draw data. Each study sends fictional applicants to actual vacancies and uses the standard outcome variable: whether the vacancy positively responds to the application by phone or e-mail. Beyond this common framework,

[10] American Economic Review, Journal of Economic Perspectives, all four American Economic Journals, Econometrica, Review of Economic Studies, Journal of Political Economy, Quarterly Journal of Economics, Review of Economics and Statistics, Journal of Labor Economics, and Journal of Human Resources.

individual studies can differ in minor ways. The first row describes the Bertrand and Mullainathan (2004) study of racial discrimination in Boston and Chicago. The main treatment of interest in this study was whether the application was assigned a stereotypically black or white name. They send 2 or 4 applications to each job vacancy. As discussed above, treatment was stratified such that half of the applications to a given job received black names and half white, inducing a correlation between an application having a black name and other applications to the same job having a white name. I use nearly their entire sample, excluding 324 observations which have missing data. The remaining 4,546 job applications represent my sample for analysis. The empirical framework summarized in equations (6) through (9) requires that in each study I identify instrumental variables for the number of callbacks to other applications. Bertrand and Mullainathan (2004) assign log median neighborhood income of the listed address and a female name dummy randomly and without stratification. They randomly assign these variables conditional on city of the job (Boston or Chicago) and type of job (administrative or sales). Hence, female names and neighborhood income provide valid instruments, but I must control for city and job type dummies. I also control for the number of applications sent to ensure that I identify spillovers off variation in the instrument Z_i rather than the number of applications. Finally, as described above, I apply a LASSO estimator to equation (9) to select the instrument with the strongest relationship between the value of instrument and callbacks to that same application. In this case, the log median neighborhood income provides a stronger instrument than a female name.

The remainder of Appendix Table 1 displays similar information for the other six studies, and I discuss these details further in an appendix. Each included study sends multiple applications to each job, at least partially stratifies treatment assignment, and includes at least one valid instrument. Given these facts, I can implement my empirical strategy using 7 different datasets spanning 15 years, 4 countries, and various types of labor and housing markets.

4. Results: Measuring Spillovers in Existing Experiments

4.1. An Extended Example with Data from One Labor Market Experiment

I find evidence of positive job search externalities in the classic Bertrand and Mullainathan (2004) study. I test for whether randomly assigning an applicant to have an address in a high income neighborhood helps not only the applicant listing the address but also other applicants to the same job. Table 1 shows the empirical results.

Recall that a valid instrument must be valued by employers, exogenous, and not stratified. Column (1) of Table 1 tests whether log median neighborhood income of the applicant is a strong instrument, i.e. whether employers value it by calling back applicants with high income addresses more frequently. I estimate the first-stage relationship between the callback rate to other applicants to the same job and average neighborhood income of other applicants. The positive and statistically significant coefficient of 0.048 indicates that doubling the neighborhood income[11] of all other applicants increases the callback rate to those same applicants by 4.8 percentage points.

I also confirm that that the neighborhood income instrument appears to be randomly assigned. Column (2) demonstrates that income of the job applicant's neighborhood does not correlate with neighborhood income of the job location. Bertrand and Mullainathan (2004) set addresses of applicants randomly. A 1 percent increase in income in the employer's neighborhood is associated with a statistically insignificant 0.005% decrease in income in the applicant's neighborhood. In a non-experimental sample, applicants would tend to apply to jobs in neighborhoods similar to their own, generating a positive coefficient, but random assignment of applicant addresses eliminates this selection problem.

[11] Because of residential sorting, neighborhood median incomes have a very wide range. In Bertrand and Mullainathan (2004) median household income ranges from $7,000 to $49,000 at the 1st and 99th percentiles of neighborhoods, respectively. Doubling neighborhood income is a large change but within the sample, for instance moving from the median of $25,000 to the 99th percentile of $49,000.

The experiment also does not stratify the random assignment of addresses by job. Column (3) demonstrates that neighborhood income of one applicant does not correlate with neighborhood income of other applicants to the same job. This fact differs from how the minority name treatment is determined. Bertrand and Mullainathan (2004) assign names to applications randomly, but they stratify this process by job vacancy so that an application with a black name competes with 1 black name while an application with a white name competes with 2 black names. Column (4) confirms this fact, demonstrating that the proportion of experimental applicants to the same job with a black name is 39 percentage points lower when the application of interest receives a black name. Unlike the original black name treatment, the experiment assigns neighborhood income of the applicant both randomly and without stratification, providing a valid instrument with which to test for spillovers across applications.

In columns (5) and (6) of Table 1, I use the neighborhood income instrument to identify job search externalities. Column (5) estimates the reduced form relationship, testing whether callbacks to a given applicant depend on the neighborhood income randomly assigned to other applicants to the same job. The coefficient of 0.046 indicates that doubling the neighborhood income of all other applicants to the same job would increase the current applicant's callback rate by 4.6 percentage points. The measured effect is large, positive, and statistically significant at the 5% level. Column (6) shows similar results using an IV framework with column (1) serving as the first stage. The IV estimation yields a positive and statistically significant coefficient of 0.95 for the spillovers coefficient. This value indicates that if the experiment increases the quality of other applicants such that they all receive one additional callback, the present applicant receives 0.95 more callbacks. This roughly 1-to-1 increase indicates very large spillovers; improving the quality of one applicant helps other applicants to the same job nearly as much as the applicant actually listing the

characteristic. Overall, I find that an applicant's callback rate responds significantly when the experiment randomly alters the quality of other applicants to the same job.

4.2. Spillovers in Several Studies: Relevance of the Instruments

I implement the same empirical strategy using data from 7 different experiments. Testing for spillovers in multiple studies provides two main advantages. First, spillover tests in one study sample may have low power or weak instruments. Some correspondence experiments generate limited random variation in applicant pool composition, and spillover effects may be much smaller than the main effects that drive sample size calculations. Pooling results across several studies increases precision. Second, applying the same identification strategy to many settings tests whether positive spillovers hold relevance for correspondence experiments in general rather than just one example.

I first test for the strength of the selected instruments pooling across the 7 datasets. In the present context, weak instruments could be a major concern. In a study with a weak instrument, the IV model in equations (7) and (8) may provide an upward biased estimate of the extent of spillovers.[12] A strong first stage will be particularly important in a context with positive spillovers. Table 2 indicates a strong first stage. The first column of Table 2 pools the first stage relationship across all 7 studies, measuring the relationship between callbacks to other applicants and the average of the instrument for those applicants. Applicants receive 2.4 percentage points more callbacks if an experiment randomly assigns them a quality characteristic one standard deviation above the mean. This coefficient is statistically significant at the 1% level indicating that employers respond as expected to a positive quality attribute. The first stage F-statistic of 77.4 easily exceeds the rule-of-thumb cutoff of 10, indicating a strong instrument. The instruments I use also match

[12] IV with weak instruments is biased toward the OLS estimates (Angrist and Pischke, 2009). OLS applied to equation (8) yields positive spillovers when call backs are positively correlated across applicants to the same job.

those the literature identifies as theoretically and empirically important.[13] One possible weak instruments concern remains. I construct the instrument by selecting the strongest of several candidate instruments. If the set of candidates is large enough, selecting empirically strong instruments could mechanically generate strong first stage F-statistics from random noise. In an appendix, I demonstrate with simulations that the variable selection procedure I use does not mechanically generate strong instruments in the present context. Overall, the quality instrument passes all standard weak instrument tests in the pooled sample.

Comparing the strength of the instrument in the pooled sample versus individual studies demonstrates the advantage of pooling into a larger sample. The remaining columns of Table 2 show the first stage for individual studies. I have clearly strong instruments in the case of Eriksson and Rooth (2014) and the housing experiment. In both cases F-statistics exceed 10. The remaining studies may have weak instruments when analyzed in isolation. F-statistics for these studies are all above 4 but below 10. Thus, spillovers measured in one particular study could be subject to bias from weak instruments and should be interpreted with care. However, pooling several studies increases statistical power, allowing me to construct a strong instrument and test for spillovers in the full sample.

4.3. Spillovers in Several Studies: Validity of the Instruments

Next, I test for the exogeneity of the instruments using baseline balance tests. If randomly assigned, the value of the instrument for other applicants to the same job should not correlate with other applicant or firm characteristics. I focus on whether the instrument correlates with the applicant having a minority name because several studies include this variable, allowing for

[13] A non-exhaustive list of studies using these instruments as a fundamental part of a correspondence study: age (Lahey, 2008; Neumark, et. al. 2015); credit/smoking (Ewens, et. al. 2014); ethnic names (Bertrand and Mullainathan, 2004; Booth, et. al. 2011); extroversion (Carlsson, et. al. 2014); neighborhood income (Bertrand and Mullainathan, 2004; Phillips, 2015).

comparison across studies. Table 3 shows the results of regressing a minority name dummy on the mean value of the instrument for other applicants to the same job. For the Bertrand and Mullainathan (2004) study, the coefficient of -0.001 indicates that when the experiment doubles the neighborhood income of other applicants to the same job, the probability that the applicant lists a black name decreases by only 0.1 percentage points, a difference which is very small and statistically insignificant. I find very similar results in data from Eriksson and Rooth (2014), Phillips (2015), and the housing study.[14] I only use instruments described by the original papers as randomly assigned, and baseline balance tests confirm that these instruments are in fact exogenous.

4.4. Spillovers in Several Studies: Measuring Spillovers

Averaging across all 7 studies, I find evidence of positive spillovers. I test whether callback rates respond to changing a quality characteristic for other experimental applicants to the same job. Table 4 shows the reduced form estimates. The first column pools over all 7 studies. Applicants benefit from reporting higher quality characteristics on their own applications. Increasing the quality of the instrument by 1 standard deviation raises callback rates by 2.0 percentage points. One applicant's quality then spills onto other applicants to the same job. Raising the quality of other applicants to the same vacancy increases the present applicant's callback rate by 0.7 percentage points. The positive spillover effect is large, about 1/3 of the direct effect on the original applicant, and statistically significant at the 1% level.

The pooled findings reflect the results of individual studies. As before, I find in the Bertrand and Mullainathan (2004) data that doubling neighborhood income of other applicants increases the present applicant's callback rate by 4.6 percentage points. The Eriksson and Rooth

[14] I cannot apply this test to the three remaining studies because they either do not include variation in name ethnicity (Lahey, 2008) or because the instrument I happen to select is only exogenous conditional on name ethnicity (Booth, et. al. 2011; Oreopoulos, 2011). However, in these three cases I use an instrument based on the original treatment variable of the study. I can use functions of the treatment variables as instruments in these three studies because they do not fully stratify treatment (see Appendix). The original studies clearly establish exogeneity of their treatment variables, and I am able to verify this using other baseline characteristics in the data. Results available on request.

(2014) study shows the most similar results. An applicant benefits from being extroverted. Extroverted applicants receive 3.3 percentage points more callbacks. The spillovers coefficient is also positive. If all other experimental applicants list themselves as being extroverted, the original applicant receives 4.5 percentage points more callbacks. This spillover effect is statistically significant at the 5% level. Four other studies show statistically insignificant coefficients but with point estimates indicating positive and reasonably large spillovers. In Lahey (2008), Oreopoulos (2011), Phillips (2015), and the housing study the own applicant and other applicant coefficients have the same sign, indicating positive though statistically insignificant point estimates for spillovers. Many statistically insignificant but economically meaningful estimates are not surprising given limited statistical power at the level of the individual study. Booth, et. al. (2011) provides the one exception to positive spillovers with own applicant and other applicant coefficients of opposite sign, though the estimated spillovers are very small. Altogether, the reduced form framework yields point estimates indicating positive spillovers in 6 of 7 studies with 2 of 7 showing statistically significant positive spillovers at the level of the individual study. Pooling results from individual studies provides strong evidence of positive spillovers on average.

I also estimate spillovers using instrumental variables, which allows me to generate comparable estimates across studies. Table 5 shows the results of implementing the IV framework of equations (7) and (8). I instrument the callback rate to other applicants with the average of the quality variable for other applicants. Pooling across all 7 studies yields a coefficient of 0.31. Improving the quality of other applicants such that they move from none getting callbacks to all getting callbacks leads the original applicant to receive 0.31 callbacks. The positive coefficient indicates positive spillovers, and the magnitude of 0.31 indicates spillovers about one third of the direct effect on the applicant actually listing the characteristic. This spillovers coefficient has a standard error of 0.09, and a 95% confidence interval from 0.13 to 0.49. In the IV estimation,

Bertrand and Mullainathan (2004), Eriksson and Rooth (2014), and Phillips (2015) show larger and statistically significant coefficients. Data from Lahey (2008), Oreopoulos (2011), and the housing study generate positive but smaller and statistically insignificant estimates. Meanwhile the results from Booth, et. al. (2011) provide a negative but small coefficient. In the full sample, spillovers are large and positive.

These results provide the most pertinent information on spillovers for the literature on racial discrimination. I find spillovers using a wide range of variables including age, extroversion, name, and neighborhood income. To gain sufficient statistical power, I average spillovers across these different contexts. If several existing experiments provided data in which race was independently assigned, I could apply the same identification strategy to measure applicant pool effects specific to race. However, the vast majority of existing studies of racial discrimination use stratified designs. In theory, a researcher could conduct a new experiment independently assigning race, but spillover effects tend to be smaller than direct effects, requiring large samples to detect. Random independent assignment of treatment also generates far less variation in group composition than individual characteristics, which makes standard errors for spillover effects large. A housing study using the methodology I employ above to measure racial spillovers would require about 100,000 fictional applications.[15] An alternative design could indirectly measure spillovers by using only applications with black names and randomly selecting vacancies to receive either 1 or 4 applications. Such a study could not separately identify spillovers from quantity versus quality of

[15] The main effect of race in the housing study is -0.053 (see below), leading to an expected spillover effect one-third of that magnitude: -0.017. The actual data provide an estimate of the standard error: restricting the housing experiment to the portion using independent assignment of treatment and 4 applications per landlord, there are 456 observations with a standard error on the spillover effect of 0.089. Thus, the required sample is $456 * \left(\frac{2.8*0.089}{0.017}\right)^2 = 97{,}986$. A carefully constructed experiment might maximize variation in the composition of the applicant pool rather than independently assigning treatment. Simulations suggest such a design can reduce the required sample size to 76,266.

other applicants but would require a sample closer to 20,000.[16] Even that value exceeds sample sizes in nearly all correspondence studies, and such an experiment is beyond the scope of the present study. A cautious experimenter would thus be well-served by judging the importance of spillovers using currently available data. In 7 experiments spanning 15 years, 4 countries, and various types of labor and housing markets, I find evidence of large, positive spillovers on average.

4.5. Applicant Pool Size and Plausible Magnitudes of Spillovers

Could a few experimental applicants affect real-world applicant pools enough to generate spillovers? Employers likely respond to the quality of the entire pool, but I can only observe and measure the average quality of a subset of fictional applicants. More generally, the size of the applicant pool could matter if a large pool dilutes the effect of any one applicant on another. The theory appendix formalizes this intuition. In a context with positive spillovers driven by employer statistical learning, appendix equation (A.4) indicates that the spillover coefficient will decrease as the pool of non-experimental applicants increases. Hence, the size of applicant pools may matter.

Existing studies indicate that applicant pools for relevant jobs range from a couple dozen to a couple hundred applicants. Correspondence experiments use a wide variety of online job sites, and representative statistics on the number of applicants on these sites are difficult to obtain. Some studies suggest quite small applicant pools. Horton (forthcoming) reports that a sample of computer technical jobs posted on the online labor market oDesk receive only 17 applicants on average. Hoffman et. al. (2015) find a hiring rate of 19% for applications to a set of low-skilled service jobs. The inverse of the hiring rate, i.e. 5 applicants per job in this case, provides a rough measure of the applicant pool. Behaghel et. al. (2015) finds that jobs hiring from French employment agencies

[16] This study moves applicant pool composition half as much, comparing adding black applicants to nothing rather than black to white. So, I would expect spillovers 1/6 of the main effect of race in the housing study, -0.053/6=0.008. If I restrict the housing experiment to the portion with either 1 or 4 applications per landlord, there are 1,377 observations. A regression on a number of applications dummy yields a standard error of 0.011. Thus, the required sample is $1377 * \left(\frac{2.8 * 0.011}{0.008}\right)^2 = 20{,}411$.

interview at a rate of 0.10 per application and hire at a rate of 0.02, implying about 50 applicants per job. The data for the present study appears similar; applicants receive a positive response from employers with probability 0.15. Hence, 50 applicants provides a useful benchmark.

In the theoretical model, a handful of fictional applicants can generate meaningful spillovers for jobs with reasonable applicant pools. Consider appendix equation (A.5). With a set of plausible parameters,[17] this equation implies a direct effect of the applicant's own characteristics, β, at about 0.062. With a pool of 50 applicants, the model implies spillovers onto other applicants of 0.018, or 29% of the direct effect. Spillovers range from 36% to 13% of the direct effect as the applicant pool increases from 25 to 200 'real' applicants. A simple model can rationalize spillovers similar to those observed empirically.

5. The Magnitude of Bias from Spillovers

5.1. Ideal Experiments to Measure Bias

Given the presence of spillovers, simple differences in callback rates calculated in stratified experiments will confound direct discrimination with changes in the applicant pool. Measuring the magnitude of this bias directly will prove difficult. Bias could be calculated by comparing racial name effects from two experiments, one with a stratified design and one without. However, such an experiment would require a prohibitively large sample. Sample sizes rise with the square of the minimum detectable effect. Detecting even a large bias equal to 25% of the measured treatment effect would multiply the required sample by 16. Also, comparing treatment effects across experiments requires interaction effects, which further increases the required sample. While direct, measuring bias through an explicit experiment will prove underpowered or cost prohibitive.

[17] Suppose that $\sigma_{(T)T} = \mu\sigma_{T\omega}/2$, $\sigma_{(T)\omega} = \mu\sigma_{T\omega}/3$, and $\mu\sigma_{T\omega} = 1$. Beyond the parameter restrictions in the text, $\sigma_{(T)\omega} < \sigma_{(T)T}$ implies that treatment correlates more with another applicant's treatment than with the other applicant's unobservables. This assumption pushes in the direction of smaller spillovers. Finally, I then pick $\sigma_T^2 = 30$ to match the direct treatment effect $\beta \approx 0.06$. I assume 4 fictional applicants sent to each job.

5.2. Approximating Bias with Simple OVB Calculations

Since additional experimentation is not practical, I quantify bias by pairing stronger assumptions with existing data. In the analysis above, I have required that the empirical specifications in equations (5) and (6) be true insofar as the data meets typical reduced form assumptions about the instrument, e.g. exogeneity. Measuring the magnitude of spillovers ***does not*** require strict assumptions about functional form. However, measuring bias ***does*** require stronger assumptions. I will not attempt to measure bias exactly based a claim of correct assumptions. Instead, consider a simple question: "Given the magnitude of measured spillovers, could a skeptical observer believe measures of discrimination in the literature include large bias?" As argued formally by Banerjee, et. al. (2016), randomized experiments prove useful precisely because they provide evidence that can convince skeptical audiences with a variety of prior beliefs. I show that an observer with a simple, plausible set of beliefs will infer large bias in estimates of discrimination when observing the magnitude of spillovers measured above.

Suppose that the econometrician runs a stratified experiment but incorrectly specifies a simple difference model as in equation (4) above. I will consider an observer who believes equations (5) and (6) from above, which include spillovers, represent the exact, true model of the world. An econometrician who estimates simple differences using data from a stratified experiment measures direct discrimination with bias equal to $\delta\beta$. An observer can measure this bias using two key assumptions embedded in a literal belief in equations (5) and (6). First, spillovers follow a simple linear form. Second, all variables spillover proportional to their direct effect. Of course, these assumptions may be false. The functional form of spillovers may not be linear, or different variables may not share the same spillover coefficient δ. Neighborhood income, age, and extroversion may generate positive spillovers at 1/3 the magnitude of the direct effect, but racially charged names may spillover at a higher fraction, lower fraction, or even negatively. Simply

detecting spillovers ***does not*** require such assumptions, but measuring bias in existing estimates ***does*** require commitment to a functional form. Thus, I consider an observer who holds simple beliefs in linear, proportional spillovers.

Under these assumptions, an observer can calculate bias using the omitted variable bias (OVB) formula. First, define a "stratification coefficient," γ_1, measuring the link between applicant pool composition and treatment:

$$\bar{T}_{(i)j} = \gamma_0 + \gamma_1 T_{ij} + v_{ij} \quad (10)$$

Then, bias can be approximated as:

$$Bias = \delta\beta(K-1) * \gamma_1 = [\delta\psi(K-1)] * \frac{\beta}{\psi} * \gamma_1 \approx [\delta\psi(K-1)] * \frac{\theta}{\psi} * \gamma_1 \quad (11)$$

The first equality is the OVB formula applied to equations (4), (5), and (10). The second equality replaces treatment spillovers, $\delta\beta(K-1)$, which cannot be estimated due to the collinearity of T_{ij} and $\bar{T}_{(i)j}$, with rescaled spillovers of the instrument. The approximation at the end results from the fact that $|\theta| < |\beta|$ in the presence of positive spillovers, which gives a conservative estimate of bias. The final expression can be computed using four regression coefficients already estimated in this paper.

5.3. Results

Table 6 applies this framework to the various datasets. The second column provides the calculations for Bertrand and Mullainathan (2004). The instrument, neighborhood income, spills over with a coefficient of 0.046. A 100% increase in the neighborhood income of other applicants increases the present applicant's callback rate by 4.6 percentage points. To obtain the spillovers coefficient for race rather than neighborhood income, we need to re-scale according to the ratio of the direct effects, $\frac{\theta}{\psi}$, as in equation (11). The direct effect of living in a 'good' neighborhood, ψ, is 2.9 percentage points. The direct effect of a black-sounding name, θ, is -3.1 percentage points.

Hence, the scaling ratio is $-\frac{0.031}{0.029}$, and the spillovers coefficient for black names can be approximated by $-\frac{0.031}{0.029} * 0.046 = -0.049$. The stratification coefficient was already estimated in the fourth column of Table 1. It is -0.39. This value is between $-\frac{1}{3}$ and -1 because they send either 2 or 4 applications to each job. Finally, the omitted variable bias formula calculates bias as:

$$Bias \approx [\delta\psi(K-1)] * \frac{\theta}{\psi} * \gamma_1 = 0.046 * \left(-\frac{0.031}{0.029}\right) * -0.39 = 0.019$$

This coefficient implies that a simple difference in callback rates underestimates discrimination against black names by 1.9 percentage points, or 38% relative to the true effect of -0.050.

Approximating bias in all studies and pooling the estimates, I find evidence consistent with large biases on average. The first column of Table 6 reports the pooled results. Studies of stereotypically minority names report an average racial name penalty of 6 percentage points. These values underestimate discrimination by 2 percentage points, on average, leading to average bias of 19%. This average value includes significant heterogeneity with 2 studies recording bias near 40%, 1 study showing bias near the average, and 3 studies indicating small biases. Calculating these values requires parametric assumptions that may fail. But a reasonable observer with a simple model of the world may expect large bias in the presence of spillovers.

5.4. Optimal Experimental Design: Simulating the Bias-Variance Tradeoff

The present results imply a tradeoff for the experimenter who wishes to estimate the direct treatment effect of an applicant characteristic separate from changes in applicant pool composition. Experimenters who use stratified designs to estimate the direct effect parameter, β, will instead measure a combination of direct and spillover effects, $\beta - \delta\beta$, and thus obtain a biased estimate. However, stratified designs are popular because they increase precision by ensuring balance of vacancy characteristics across treatment and control groups. Suppose the experimenter wishes to minimize mean squared error (MSE) in the process of estimating β:

$$MSE = E\left[\left(\hat{\beta} - \beta\right)^2\right] = \left(E[\hat{\beta}] - \beta\right)^2 + E\left[\left(\hat{\beta} - \beta\right)^2\right] = Bias(\hat{\beta})^2 + Var(\hat{\beta})$$

Stratified designs increase MSE due to bias but decrease MSE via efficiency gains. Optimal experimental design depends on the magnitudes of these competing effects.

I use the bias estimates described above and Monte Carlo simulations to test whether stratified designs generate sufficient efficiency gains to outweigh any bias. An appendix provides the technical details. Figure 1 shows the results, plotting the estimated ratio of mean squared error for non-stratified versus stratified designs. In a sample of 1,237 vacancies equal to the original sample size in Bertrand and Mullainathan (2004), a non-stratified design generates MSE roughly 1/6 of the stratified design. The non-stratified design similarly outperforms a stratified design averaging over all available studies. Stratification does provide greater efficiency benefits in small samples, and stratified designs become MSE-preferred for studies sampling fewer than 100 vacancies. However, reasonably powered studies, including all of the datasets I use, use samples much larger than 100 vacancies. Stratified designs may still be preferred for small-scale initial pilots and small sample audit studies using real people as testers, but researchers should expect lower MSE from non-stratified designs at the sample sizes used in reasonably-powered, modern correspondence experiments. The bias generated by stratifying appears to outweigh any potential efficiency gains.

6. A Toolbox for Designing Experiments in the Presence of Spillovers

The results above suggest that correspondence experiments using stratified treatment can generate large spillovers that bias estimates of discrimination. The cautious researcher will thus consider alternative strategies for experimental design. Moving from most to least robust in addressing spillovers, these strategies include sending one application to each vacancy or multiple applications without stratifying, using a non-stratified sub-sample to bound bias, and ex-post approximating bias using the OVB formula.

6.1. Sending One Application per Vacancy or Multiple Applications without Stratifying

The simplest and most robust response to spillovers is to send one application per vacancy or send multiple applications while independently assigning the treatment to each application. As discussed above, these strategies alleviate the bias from spillovers identified in this paper. Both of these options will impose a loss of statistical precision, necessitating a larger sample size. Sending only one application per vacancy will also incur a large time cost. Hence, switching to a non-stratified design with multiple applications per vacancy will likely be the most cost-effective choice if the only concern is spillovers across applicants. However, sending multiple applications to a job can cause other problems, such as increasing the likelihood of detection by the employer (e.g. Weichselbaumer, 2014), which should be considered depending on the context.

6.2. Bounding Discrimination by Measuring Spillovers in a Sub-Sample Experiment

In some cases, experimenters may already have data from a stratified experiment which would be costly to replicate. In this context, the experimenter can measure spillovers rigorously by adding a sub-sample in which some vacancies receive stratified treatment assignment and others receive independent treatment assignment (or one application). The researcher can then bound estimates of discrimination in the larger stratified sample with this measure of spillovers. This bounding exercise will typically require a large sample to be informative.

Returning to the notation of equations (4) and (5), suppose the researcher has an estimate of discrimination from a stratified experiment, $\beta(\widehat{1-\delta})$, that potentially confounds direct discrimination and spillovers from applicant pool composition. However, the researcher has also conducted an auxiliary experiment on a separate sample within which she randomly assigns vacancies to stratified or independent treatments. The researcher then estimates:

$$Y_{ij} = \alpha + \theta T_{ij} + \xi I_j + \omega I_j * T_{ij} + v_{ij} \quad (12)$$

where I_j is a dummy for vacancies with treatment assigned independently, i.e. not stratified. The interaction coefficient $\widehat{\omega}$ consistently estimates the difference between discrimination in the non-stratified and stratified vacancies, which measures spillovers: $\widehat{\omega} \sim N\big(\beta\delta, Var(\widehat{\omega})\big)$. Adding this estimate to the full stratified sample estimate nets out spillovers, yielding a consistent estimate of discrimination only:

$$\widehat{\beta(1-\delta)} + \widehat{\omega} \sim N\left(\beta, Var\big(\widehat{\beta(1-\delta)}\big) + Var(\widehat{\omega}) + 2Cov(\beta(1-\delta), \widehat{\omega})\right)$$

Consider how this method would work in the context of the housing experiment used in the analysis above. As described in an appendix, I stratify randomization of a name treatment for some apartments but not others. I can then measure and compare racial discrimination in these two sub-experiments. The first three columns of Table 7 show these results. The sample using stratified randomization measures a statistically significant 6.7 percentage point decrease in positive response rates to messages with black names. However, the non-stratified sample measures an effect of 3.1 percentage points that is statistically insignificant and roughly half as large. As shown in the final column, the difference between these two estimates in not statistically significant due to limited power.

This comparison of experimental designs can be used to bound treatment effects in a larger stratified experiment. Suppose that the experimenter has already conducted an experiment with 20,000 applications that stratifies a racial name treatment by vacancy. Suppose this experiment finds that applicants with black-sounding names receive 7.0 percentage points fewer callbacks with a standard error of 0.6 percentage points. Concerned about spillovers, the experimenter conducts a post-test on a new sample that randomly assigns some vacancies to stratified treatment and others to independent treatment. This post-test yields the analysis shown in Table 7. The interaction term in column (3) measures bias from spillovers to be 3.7 percentage points. Adjusting for these spillovers, applicants with black names receive 3.3 percentage points fewer callbacks. This

estimate has a standard error of $\sqrt{0.006^2 + 0.039^2 + 2 * -0.00065} = 0.016$ or 1.6 percentage points.[18] Thus, the researcher would conclude that discrimination still exists, even accounting for spillovers. The 95% confidence interval for the gap between white and black names would be $3.3 \pm 1.96 * 1.6$. Such bounding exercises will be most informative for experiments with large samples, but they do provide a less drastic response than abandoning data from a large-scale, stratified experiment.

6.3. Ex-Post Approximating Bias Using the OVB Formula

If the researcher cannot implement a subsample with non-stratified treatment, he may use the method from the present paper exploiting a quality instrument and the omitted variable bias formula. The method can be implemented exactly as in the present study. It requires data on some ancillary characteristic from each application that is randomly assigned, not stratified, and valued by employers. Spillovers can be measured as in equation (6). If this method detects spillovers, then the researcher can adjust estimates for bias as in equation (11).

However, adjusting for bias using instrumental variables requires additional, untestable assumptions that may not hold in any given sample. Consider the housing experiment. The OVB formula implies that standard experiments underestimate treatment effects in the presence of positive spillovers, as shown in Table 6. However, the first three columns of Table 7 experimentally measure bias from spillovers and find the opposite result; the stratified experiment appears to overestimate discrimination. These conflicting results could reflect noise caused by low power. However, it is possible that at least one assumption of equations (5) and (6) are violated. This model of the world assumes that the treatment variable and the instrumental variable spillover

[18] The first term is the squared standard error of the treatment effect in the larger experiment. The second term of the variance is the squared standard error of the interaction term in column (3) of Table 7, and the final term is the covariance of the interaction coefficient and the treatment coefficient, which is computed as part of the regression in column (3).

in the same direction and proportional to their direct effect. Since race is not fully stratified in the housing experiment, I can test this assumption directly, seeing whether spillovers by race are similar to those of the chosen instrument. The final column of Table 7 provides some evidence to this end, showing a reduced form regression of a positive response dummy on the low credit/smoking instrument and racial name variables. As above, having a low credit/smoking message on one's application leads to a large reduction in callback rates of 27 percentage points. Having that negative signal in the e-mail of all other applicants to the same apartment also hurts the present applicant, reducing callback rates by 4 percentage points. In terms of point estimates, race spills over in the opposite direction. Listing a black-sounding name on one's own application reduces callbacks by 3.9 percentage points, but having all other experimental applicants list a black name increases the callback rate by 3.9 percentage points. These results are not statistically significant, and real differences cannot be separated from statistical noise. However, the point estimates provide an illustration of how adjusting for spillovers using a parametric formulation could fail. In general, racially-charged names may spillover differently from available instruments. Thus, a cautious experimenter would need to take care when adjusting measured discrimination ex-post based on spillovers measured with an instrument. In most cases, a sound experimental design will prove better than this technical fix.

7. Conclusion

I document a simple but surprisingly common flaw in interpreting correspondence studies of discrimination. Most studies use gaps in callback rates to measure how identical people will experience different levels of success in the same application process if they differ on only one attribute, e.g. race. However, common experimental designs alter the composition of the applicant pool differently for applicants with and without the attribute of interest. Matched pair and stratified experimental designs assign a fixed proportion of applicants from each vacancy to the treatment

group. In such designs, differences in callback rates measure how employers will respond if we change one applicant's name to signal a minority group *and* change a second competing applicant's name to stop signaling the minority group. If an employer's response to one applicant depends on other applicants' attributes, stratified designs will confound the direct effect of a trait with changes in the applicant pool.

Empirically, spillovers across applicants commonly occur. I detect spillovers by exploiting random variation in the quality of the applicant pool generated by the designs of several existing correspondence experiments. I test whether one applicant's response rate changes when the experiment randomly assigns another applicant to have higher quality. Across seven studies, I find strong evidence of positive spillovers; on average, giving one applicant a high quality attribute raises the callback rate to other applicants to the same job by 31% of the effect on the original applicant. When I use one simple parameterization to remove applicant pool composition effects, measured discrimination based on the applicant's name increases by 19%.

The sensitivity of measured discrimination to spillovers complicates efforts to interpret existing correspondence studies. Most researchers interpret correspondence experiments as direct discrimination against an individual's attribute rather than the effect of changing the composition of the applicant pool. A clearer interpretation would recognize that differences in callback rates combine the response of an employer to the applicant's credentials and the credentials of the applicant pool overall. Spillovers also make comparisons between studies challenging. For example, Bertrand and Mullainathan (2004) and Booth, et. al. (2011) both study racial discrimination using employer responses to names. As shown above, the former study fully stratifies treatment, shows signs of significant spillovers, and thus combines the direct effect of discrimination with applicant pool composition effects. The latter study only partially stratifies treatment, shows no signs of spillovers, and thus measures only the direct effect of discrimination.

A researcher wishing to summarize the literature or compare studies will face a significant challenge because many apparently similar studies do not estimate the same parameter.

The present problem has a simple solution. Researchers designing correspondence experiments can abandon stratified experimental designs, including the matched pairs design. Stratifying does create noticeable efficiency gains, but a risk of very large bias makes non-stratified designs better on average. Using simple Monte Carlo simulations, I find that the bias caused by stratified experimental designs in the presence of spillovers outweighs any efficiency gains. While efficiency gains become more important in very small samples, I estimate that studies applying to as few as 100 vacancies will be better served by non-stratified designs. I also find the bias caused by stratification to be unpredictable. Overall, a researcher wishing to estimate how employers respond to two otherwise identical individuals engaging in an identical job search should avoid the stratified/matched pairs design.

Several feasible options remain. The experimenter can still send multiple applications to each vacancy while independently assigning treatment for each applicant (e.g. Kroft, et. al. 2013) or simply send one application to each vacancy (e.g. Ewens, et. al. 2014). Alternatively, researchers can test for and bound spillover bias in an existing dataset by implementing a smaller sample post-test that compares estimates from stratified and non-stratified designs. These options will generally perform better than ex-post bias adjustment based on simple parametric assumptions that may not generally hold. Overall, spillovers across applicants significantly complicate the interpretation and comparison of correspondence studies, but such problems can be fixed in future studies with straightforward changes to experimental design.

References

Albrecht, James, Gerard J. Van den Berg, and Susan Vroman (2009). "The aggregate labor market effects of the Swedish knowledge lift program." *Review of Economic Dynamics* 12, no. 1: 129-146.

Angrist, J. and J.S. Pischke (2009) *Mostly Harmless Econometrics*. Princeton University Press.

Arceo-Gomez, E. O., and R.M. Campos-Vazquez (2014) 'Race and Marriage in the Labor Market: A Discrimination Correspondence Study in a Developing Country.' *The American Economic Review: Papers and Proceedings, 104*(5).

Baert, S., Cockx, B., Gheyle, N., and Vandamme, C. (2015) "Is there less discrimination in occupations where recruitment is difficult?" *ILR Review*.

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*.

Banerjee, A., Chassang, S., and Snowberg, E. (2016) "Decision Theoretic Approaches to Experiment Design and External Validity." *NBER Working Papers,* No. 22167.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *The Journal of Economic Perspectives*, 29-50.

Behaghel, L., Crépon, B., & Barbanchon, T. L. (2015). Unintended effects of anonymous resumes. *American Economic Journal: Applied Economics, 7*(3), 1-27.

Bertrand, M. and S. Mullainathan (2004) 'Are Emily and Greg More Employable Than Lakisha and Jamal." *American Economic Review,* 94(4).

Blundell, Richard, Monica Costa Dias, Costas Meghir, and John Reenen (2004) "Evaluating the employment impact of a mandatory job search program." *Journal of the European Economic Association* 2, no. 4: 569-606.

Booth, A. L., Leigh, A., & Varganova, E. (2012). Does ethnic discrimination vary across minority groups? Evidence from a field experiment. *Oxford Bulletin of Economics and Statistics, 74*(4), 547-573.

Carlsson, M., Fumarco, L., & Rooth, D. O. (2014). Does the design of correspondence studies influence the measurement of discrimination?. *IZA Journal of Migration, 3*(1), 1-17.

Cox D.R. (1958) *Planning of Experiments*. Wiley; New York.

Cragg, J.G. and Donald, S.G. (1993) Testing Identfiability and Specification in Instrumental Variables Models. *Econometric Theory,* Vol. 9, pp. 222-240.

Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., & Zamora, P. (2013). "Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment." *The Quarterly Journal of Economics, 128*(2), 531-580.

Dahlberg, Matz, and Anders Forslund (2005) "Direct Displacement Effects of Labour Market Programmes." *The Scandinavian Journal of Economics* 107, no. 3: 475-494.

Deming, D. J., Yuchtman, N., Abulafi, A., Goldin, C., & Katz, L. F. (2016). The value of postsecondary credentials in the labor market: An experimental study. *The American Economic Review*, *106*(3), 778-806.

Eriksson, S. & Rooth, D. (2014). Do employers use unemployment as a sorting criterion when hiring? Evidence from a field experiment. *The American Economic Review*, *104*(3), 1014-1039.

Ewens, M., Tomlin, B., & Wang, L. C. (2014). Statistical discrimination or prejudice? A large sample field experiment. *Review of Economics and Statistics*, *96*(1), 119-134.

Ferracci, M., Jolivet, G., & van den Berg, G. J. (2014). Evidence of treatment spillovers within markets. *Review of Economics and Statistics*, *96*(5), 812-823.

Gautier, Pieter, Paul Muller, Bas van der Klaauw, Michael Rosholm, Michael Svarer (2012) "Estimating Eqilibrium Effects of Job Search Assistance." Institute for the Study of Labor. IZA Discussion Paper 6748.

Hanson, A., & Hawley, Z. (2011). Do landlords discriminate in the rental housing market? Evidence from an internet field experiment in US cities. *Journal of Urban Economics*, *70*(2).

Heckman, J. (1998) "Detecting Discrimination." *Journal of Economic Perspectives*, 12(2).

Hoffman, M., Kahn, L. B., & Li, D. (2015). "Discretion in hiring." National Bureau of Economic Research Working Paper No. w21709.

Horton, J. (Forthcoming) "The Effects of Algorithmic Labor Market Recommendations: Evidence from a Field Experiment." *Journal of Labor Economics*.

Hosios, A. J. (1990). "On the efficiency of matching and related models of search and unemployment." *The Review of Economic Studies*, *57*(2), 279-298.

Lahey, J. N. (2008). Age, Women, and Hiring An Experimental Study. *Journal of Human Resources*, *43*(1), 30-56.

Lalive, R., Landais, C., & Zweimüller, J. (2015). Market externalities of large unemployment insurance extension programs. *The American Economic Review*, *105*(12), 3564-3596.

Manski, Charles F. (1993) "Identification of endogenous social effects: The reflection problem." *The Review of Economic Studies* 60, no. 3: 531-542.

Neumark, D. (2012) "Detecting Discrimination in Audit and Correspondence Studies," *Journal of Human Resources*, 47(4).

Neumark, D., I. Burn, and P. Button (2015). "Is It Harder for Older Workers to Find Jobs? New and Improved Evidence from a Field Experiment." National Bureau of Economic Research.Working Paper No. 21669.

Neumark, D. and J. Rich (2016) Do Field Experiments on Labor and Housing Markets Overstate Discrimination? Re-examination of the Evidence. National Bureau of Economic Research Working Paper No. 22278.

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., Western, B., & Bonikowski, B. (2009). Discrimination in a low-wage labor market a field experiment. *American Sociological Review*, *74*(5), 777-799.

Pallais, A. (2014) "Inefficient Hiring in Entry-Level Labor Markets." *American Economic Review*, *104*(11), 3565-99.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review*, *62*(4), 659-661.

Phillips, D. (2015) "Neighborhood Affluence of Long Commutes: Using a Correspondence Experiment to Test Why Employers Discriminate Against Applicants from Poor Neighborhoods." *UC Davis Center for Poverty Research Working Paper*.

Pissarides, C. (2000) *Equilibrium Unemployment Theory, 2nd Edition.* MIT Press: Cambridge.

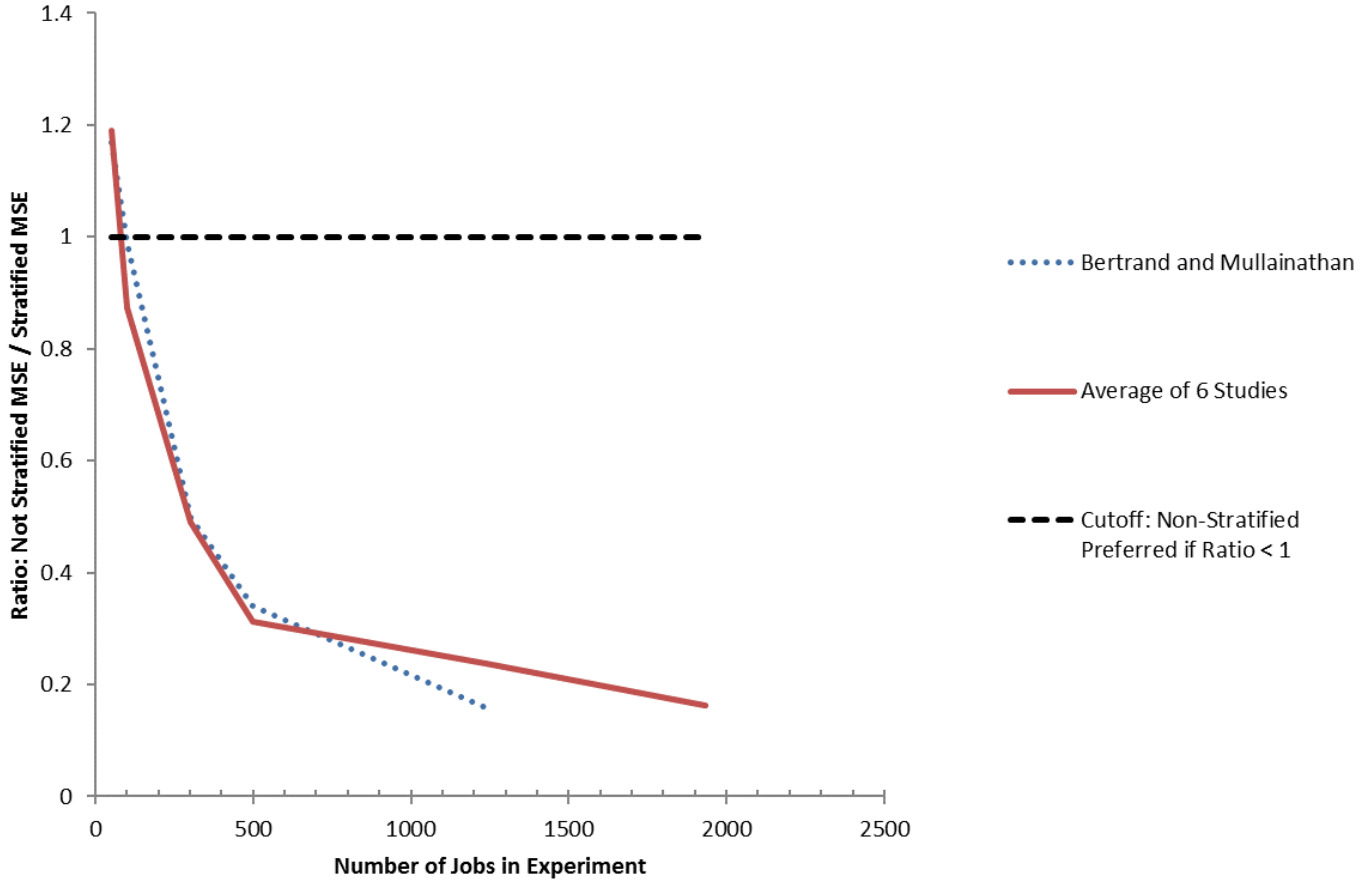
Riach, P.A. and Rich, J. (2002) Field experiments of discrimination in the market place. *The Economic Journal*, *112*(483).

Stock, J.H. and Yogo, M. (2005) "Testing for Weak Instruments in Linear IV Regression." In D.W.K. Andrews and J.H. Stock, eds. *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge: Cambridge U Press.

Weichselbaumer, D. (2015) "Testing for Discrimination against Lesbians of Different Marital Status: A Field Experiment." *Industrial Relations: A Journal of Economy and Society*, 54: 131–161.

Yinger, J. (1995). *Closed doors, opportunities lost: The continuing costs of housing discrimination*. Russell Sage Foundation.

Figure 1: Mean Squared Error in Non-Stratified vs. Stratified Designs



Notes: See appendix for details on MSE calculations.

Table 1. Spillovers in Bertrand and Mullainathan (2004)

	(1)	(2)	(3)	(4)	(5)	(6)
	First Stage	Baseline Balance	Stratification Test	Stratification Test	Spillovers (Reduced Form)	Spillovers (IV)
Dependent Variable:	Fraction of Other Applications Receiving Callbacks	Log Median Income of Applicant's Neighborhood	Log Median Income of Applicant's Neighborhood	Fraction of Other Applications with Black Names	Callback Dummy	Callback Dummy
Log Median Income - Mean Over Other Applications	0.048**	--	-0.005	--	0.046**	--
	(0.021)		(0.033)		(0.023)	
Log Median Income of Employer's Neighborhood	--	-0.005	--	--	--	--
		(0.017)				
Black name	--	--	--	-0.39***	--	--
				(0.004)		
Log Median Income	0.019**	--	--	--	0.029***	0.011
	(0.009)				(0.011)	(0.015)
Callback – Fraction of Other Applications	--	--	--	--	--	0.95***
						(0.27)
Control Variables	YES	YES	YES	YES	YES	YES
Sample size	4,546	1,760	4,546	4,546	4,546	4,546

Statistical significance at the 1, 5, and 10 percent levels is denoted by ***, **, and * respectively. Applicant controls include a city dummy, sales job dummy, and a dummy for only 2 applications sent. Standard errors are clustered at the job vacancy level. The second column has a smaller sample because some employer neighborhood income is missing.

Table 2. Spillovers in Several Correspondence Studies, First Stage

Study	Pooled	Bertrand and Mullainathan	Booth, et. al.	Eriksson and Rooth	Lahey	Oreopoulos	Phillips	Housing
Estimation Method:	OLS Callback Rate to Other Applicants	OLS Callback Rate to Other Applicants	OLS Callback Rate to Other Applicants	OLS Callback Rate to Other Applicants	OLS Callback Rate to Other Applicants	OLS Callback Rate to Other Applicants	OLS Callback Rate to Other Applicants	OLS Callback Rate to Other Applicants
Dependent Variable:								
Instrument – Others	0.024***	0.048**	0.097**	0.055***	-0.0009***	0.067**	-0.0031**	-0.30***
	(0.003)	(0.021)	(0.43)	(0.017)	(0.0003)	(0.031)	(0.0015)	(0.04)
Instrument – Own	0.004***	0.019**	-0.00	0.022**	-0.0002	0.005	-0.0008	-0.04
	(0.002)	(0.009)	(0.02)	(0.009)	(0.0003)	(0.010)	(0.0005)	(0.03)
First Stage F Statistic	76.8	5.4	4.9	11.1	7.6	4.6	4.2	72.7
Cragg-Donald Statistic	142.6	10.5	11.6	21.1	10.6	13.6	10.6	79.6
Stock-Yogo 10% Cutoff	16.4	16.38	16.38	16.38	16.38	16.38	16.38	16.38
Stock-Yogo 15% Cutoff	9.0	8.96	8.96	8.96	8.96	8.96	8.96	8.96
Control Variables	YES	YES	YES	YES	YES	YES	YES	YES
Control Group Callback Rate	0.15	0.10	0.35	0.28	0.06	0.14	0.21	0.46
Sample size	40,947	4,546	4,210	6,873	7,932	12,906	2,260	2,681

Statistical significance at the 1, 5, and 10 percent levels is denoted by ***, **, and * respectively. Applicant controls and instruments used are listed in Appendix Table 1. Pooled column includes study dummies, specifies the instrument as a z-score, and sets control variables from other studies to a value of zero. Standard errors are clustered at the job vacancy level.

Table 3. Spillovers in Several Correspondence Studies, Baseline Balance

Study	Bertrand and Mullainathan	Booth, et. al.	Eriksson and Rooth	Lahey	Oreopoulos	Phillips	Housing
Estimation Method:	OLS	--	OLS	--	--	OLS	OLS
Dependent Variable:	Minority Name Dummy	--	Minority Name Dummy	--	--	Minority Name Dummy	Minority Name Dummy
Instrument – Others	-0.001	--	0.002	--	--	-0.0014	0.001
	(0.025)		(0.011)			(0.0011)	(0.039)
Control Variables	YES	--	YES	--	--	YES	YES
Control Group Callback Rate	0.10	--	0.28	--	--	0.21	0.46
Sample size	4,546	--	6,873	--	--	2,260	2,681

Statistical significance at the 1, 5, and 10 percent levels is denoted by ***, **, and * respectively. Applicant controls and instruments used are listed in Appendix Table 1. Standard errors are clustered at the job vacancy level. I exclude Lahey (2008), Booth et al (2011) and Oreopoulos (2011) from this table for lack of a minority name variable or because the instrument I use is exogenous only conditional on the minority name variable. In all three cases, my instrument is derived from the original study's treatment variable for which exogeneity is clearly established in the original study

Table 4. Spillovers in Several Correspondence Studies, Reduced Form

Study	Pooled	Bertrand and Mullainathan	Booth, et. al.	Eriksson and Rooth	Lahey	Oreopoulos	Phillips	Housing
Estimation Method:	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS
Dependent Variable:	Callback Dummy	Callback Dummy	Callback Dummy	Callback Dummy	Callback Dummy	Callback Dummy	Callback Dummy	Callback Dummy
Instrument – Others	0.007***	0.046**	-0.003	0.045**	-0.0002	0.011	-0.0023	-0.043
	(0.003)	(0.023)	(0.044)	(0.017)	(0.0003)	(0.029)	(0.0016)	(0.041)
Instrument – Own	0.020***	0.029***	0.10***	0.033***	-0.0009***	0.057***	-0.0016**	-0.273***
	(0.002)	(0.011)	(0.02)	(0.010)	(0.0003)	(0.017)	(0.0007)	(0.029)
Others/Own	0.36	1.59	-0.03	1.46	0.11	0.19	1.44	0.16
		Joint hypothesis test of no spillovers for all studies: $F = 2.03,\ p = 0.05$						
Control Variables	YES	YES	YES	YES	YES	YES	YES	YES
Control Group Callback Rate	0.15	0.10	0.35	0.28	0.06	0.14	0.21	0.46
Sample size	40,947	4,546	4,210	6,873	7,932	12,906	2,260	2,220

Statistical significance at the 1, 5, and 10 percent levels is denoted by ***, **, and * respectively. Applicant controls and instruments used are listed in Appendix Table 1. The pooled column includes study dummies, specifies the instrument as a z-score, and sets control variables from other studies to a value of zero. Standard errors are clustered at the job vacancy level. A fully-interacted specification on the pooled sample replicates the individual study results and forms the basis of the joint test of whether the study-specific spillover coefficients in the first row are all zero.

Table 5. Spillovers in Several Correspondence Studies, 2SLS

Study	Pooled	Bertrand and Mullainathan	Booth, et. al.	Eriksson and Rooth	Lahey	Oreopoulos	Phillips	Housing
Estimation Method:	2SLS	2SLS	2SLS	2SLS	2SLS	2SLS	2SLS	2SLS
Dependent Variable:	Callback Dummy	Callback Dummy	Callback Dummy	Callback Dummy	Callback Dummy	Callback Dummy	Callback Dummy	Callback Dummy
Callback Rate to Other Applicants	0.31***	0.95***	-0.04	0.80***	0.17	0.16	0.75***	0.14
	(0.09)	(0.27)	(0.47)	(0.17)	(0.30)	(0.37)	(0.25)	(0.13)
Instrument – Own	0.019***	0.011	0.10***	0.015	-0.0009***	0.056***	-0.001	-0.27***
	(0.002)	(0.015)	(0.02)	(0.013)	(0.0003)	(0.016)	(0.001)	(0.03)
		Joint hypothesis test of no spillovers for all studies: $\chi^2 = 44.40,\ p = 0.00$						
Control Variables	YES	YES	YES	YES	YES	YES	YES	YES
Control Group Callback Rate	0.15	0.10	0.35	0.28	0.06	0.14	0.21	0.46
Sample size	40,947	4,546	4,210	6,873	7,932	12,906	2,260	2,220

Statistical significance at the 1, 5, and 10 percent levels is denoted by ***, **, and * respectively. Applicant controls and instruments used are listed in Appendix Table 1. Pooled column includes study dummies, specifies the instrument as a z-score, and sets control variables from other studies to a value of zero. Standard errors are clustered at the job vacancy level. A fully-interacted specification on the pooled sample replicates the individual study results and forms the basis of the joint test of whether the study-specific spillover coefficients in the first row are all zero.

Table 6. Approximating the Magnitude of Bias in Discrimination Estimates

Study	Pooled	Bertrand and Mullainathan	Booth, et. al.	Eriksson and Rooth	Oreopoulos	Phillips	Housing
Simple Treatment Effect - Instrument	0.072	0.029	0.100	0.033	0.057	0.002	0.273
Spillovers Coefficient - Instrument	0.024	0.046	-0.003	0.045	0.011	0.002	0.043
Simple Treatment Effect - Race	-0.060	-0.031	-0.086	-0.106	-0.032	-0.061	-0.053
Rescaled Spillovers Coefficient - Race	-0.048	-0.049	0.003	-0.145	-0.006	-0.088	-0.008
Stratification Coefficient	-0.34	-0.39	-0.27	-0.50	-0.28	-0.25	-0.35
Bias	0.020	0.019	-0.001	0.072	0.002	0.022	0.003
% Bias	-19%	-38%	1%	-41%	-5%	-26%	-5%
Sample size	40,947	4,546	4,210	6,873	12,906	2,260	2,220

I complete all analysis within studies except the first column which averages across studies weighting by the square root of the sample size. The treatment and spillover coefficients for the instrument come from Table 4. For the simple treatment effect for race, I regress a callback dummy on a minority name dummy and study-specific control variables. The re-scaled spillovers coefficient equals the instrument's spillovers coefficient multiplied by the ratio of the direct effects for race and the instrument. For the stratification coefficient, I regress the proportion of other applicants with a minority name on a minority name treatment dummy and study-specific controls. Bias equals the race spillovers coefficient multiplied by the stratification coefficient. I exclude Lahey (2008) because it does not include a minority name treatment.

Table 7. Regressions Related to Toolbox

Dataset:	Housing	Housing	Housing	Housing
Estimation Method:	OLS	OLS	OLS	OLS
Dependent Variable:	Callback Dummy	Callback Dummy	Callback Dummy	Callback Dummy
Black Name	-0.068***	-0.031	-0.068***	-0.039
	(0.026)	(0.029)	(0.026)	(0.025)
Not Stratified	--	--	-0.020	--
			(0.033)	
Black X Not Stratified	--	--	0.037	--
			(0.039)	
Black Name – Others	--	--	--	0.039
				(0.030)
Bad Credit/Smoker	--	--	--	-0.274***
				(0.029)
Bad Credit/Smoker – Others	--	--	--	-0.041
				(0.041)
Control Variables	YES	YES	YES	YES
Control Group Callback Rate	0.43	0.43	0.43	0.43
Sample	Stratified and At Least 2 Apps	Non-Stratified and At Least 2 Apps	At Least 2 Apps	At Least 2 Apps
Sample size	1,122	1,098	2,220	2,220

Statistical significance at the 1, 5, and 10 percent levels is denoted by ***, **, and * respectively. Control variables include a dummy for the second experimental phase in which more applications were sent per vacancy on average and a set of dummies for the number of applications sent. Standard errors are clustered at the apartment vacancy level.

For Online Publication Only

Appendix 1. A Simple Formal Theory of Spillovers across Applicants

Suppose that an employer observes applicants for vacancy j and decides which applicants to interview. The employer values applicant i at value P_{ij}:

$$P_{ij} = \mu T_{ij} + \omega_{ij} + u_{ij}$$

P_{ij} measures productivity of the worker as well as any preference the employer has for not hiring a particular worker, i.e. due to taste-based discrimination. This summary measure of preference depends on an observable characteristic T_{ij}. Productivity also depends on a characteristic observed by the employer but not the econometrician u_{ij} and a characteristic ω_{ij} observed by neither the employer nor the econometrician. For simplicity, suppose that u_{ij} is a random, job-applicant specific shock uncorrelated with all other variables.

The employer receives a batch of K applications and observes characteristics for all applicants. K^R of the applications come from real applicants and $K^F = K - K^R$ fictional applicants come from the experiment. Define $\boldsymbol{T_j}$ and $\boldsymbol{u_j}$ as vectors containing characteristics observed by the employer for all applicants and $\boldsymbol{\omega_j}$ as a vector of unobserved characteristics. These vectors can be split into the characteristics of person i and the characteristics of other applicants to job j, i.e. $\boldsymbol{T_j} = [T_{ij}, \boldsymbol{T_{(i)j}}]$. Given this information, the employer forms expectations regarding the value of each applicant:

$$E\left[P_{ij} \middle| \boldsymbol{T_j}, \boldsymbol{u_j}\right] = \mu T_{ij} + E\left[\omega_{ij} \middle| \boldsymbol{T_j}\right] + u_{ij} \quad (A.1)$$

The employer directly observes the applicant's own characteristics T_{ij} and u_{ij} and includes them in the evaluation. The employer also infers an expected value of the unobserved characteristic, ω_{ij}, based on observed information. For concreteness, the employer believes that $\boldsymbol{\omega_j}$ and $\boldsymbol{T_j}$ have a joint normal distribution with the following moments:

$$Var(\omega_{ij}) = \sigma_\omega^2 \quad \forall\, i$$
$$Var(T_{ij}) = \sigma_T^2 \quad \forall\, i$$
$$Cov(T_{kj}, T_{ij}) = \sigma_{(T)T} \quad \forall\, k \neq i$$
$$Cov(T_{ij}, \omega_{ij}) = \mu\sigma_{T\omega} \quad \forall\, i$$
$$Cov(T_{kj}, \omega_{ij}) = \mu\sigma_{(T)\omega} \quad \forall\, k \neq i$$

Assume all σ parameters to be strictly positive. The inclusion of μ in the last two definitions implies that a productive treatment ($\mu > 0$) such as college completion implies more productive unobservables. I will also make the mild assumptions that applicant characteristics are not perfectly correlated ($\sigma_T^2 > \sigma_{(T)T}$) and that an applicant's own characteristic more highly correlates with his own unobserved productivity than do the characteristics of other applicants ($\sigma_{T\omega} > \sigma_{(T)\omega}$).

Given this setup, standard results for the normal distribution imply:

$$E\left[\omega_{ij}\middle|\boldsymbol{T_j}\right] = cst. + \begin{bmatrix} \mu\sigma_{T\omega} \\ \mu\sigma_{(T)\omega} \\ \dots \\ \mu\sigma_{(T)\omega} \end{bmatrix}' \begin{bmatrix} \sigma_T^2 & \sigma_{(T)T} & \dots & \sigma_{(T)T} \\ \sigma_{(T)T} & \sigma_T^2 & \dots & \sigma_{(T)T} \\ \dots & \dots & \dots & \dots \\ \sigma_{(T)T} & \sigma_{(T)T} & \dots & \sigma_T^2 \end{bmatrix}^{-1} \boldsymbol{T_j}$$

It can be shown that this is equivalent to:

$$E\left[\omega_{ij}\middle|\boldsymbol{T_j}\right] = \mu\Gamma_0\Gamma_1 T_{ij} + \mu\Gamma_0\left[\frac{\sigma_{(T)\omega}}{\sigma_{T\omega}}\sigma_T^2 - \sigma_{(T)T}\right]\sum_{k\neq i} T_{kj} + cst. \quad (A.2)$$

where:

$$\Gamma_0 = \frac{\sigma_{T\omega}}{\left[(K-1)\sigma_{(T)T} + \sigma_T^2\right][\sigma_T^2 - \sigma_{(T)T}]} > 0$$

$$\Gamma_1 = \sigma_{\mathrm{T}}^2 + (\mathrm{K}-2)\sigma_{(T)T} - \frac{\sigma_{(T)\omega}}{\sigma_{T\omega}}(\mathrm{K}-1)\sigma_{(T)T} > 0$$

Combining (A.1) and (A.2), the employer forecasts productivity to be:

$$E\left[P_{ij}\middle|\boldsymbol{T_j}, \boldsymbol{u_j}\right] = \mu(1 + \Gamma_0\Gamma_1)T_{ij} + \mu\Gamma_0\left[\frac{\sigma_{(T)\omega}}{\sigma_{T\omega}}\sigma_T^2 - \sigma_{(T)T}\right]\sum_{k\neq i} T_{kj} + cst. \quad (A.3)$$

The coefficient on T_{ij} in equation (A.3) exhibits tasted-based (μ) and statistical ($\mu\Gamma_0\Gamma_1$) discrimination by the employer. Employers who see an applicant with a positive characteristic also infer that this person has more productive unobserved characteristics.

More important for the present context, the employer's evaluation of a given applicant also depends on other applicants through the coefficient on the characteristics of other applicants to the same job, $\sum_{k\neq i} T_{kj}$. Employers may look favorably on one applicant if other applicants signal productive characteristics. These positive spillovers occur when $\left[\frac{\sigma_{(T)\omega}}{\sigma_{T\omega}}\sigma_T^2 - \sigma_{(T)T}\right]$ is positive, i.e. the observed characteristics of the applicant pool provide information about another applicant's unobserved characteristics ($\sigma_{(T)\omega} > 0$) beyond what can already be known from that applicant's observable traits ($-\sigma_{(T)T}$).

Recall the example of a racial name treatment. A prejudiced employer wishes to avoid applicants who are black, indicated by ω_{ij}, but can only observe a black-sounding name treatment, T_{ij}. Suppose that job postings that attract one obviously black applicant tend to attract other black applicants with neutral names. Then, the employer will gain information about whether one applicant is black from the name of another applicant, $\sigma_{(T)\omega} > 0$. Another applicant with an obviously black name may change the employer's opinion of other applicants with observed but less informative names.

To take this model to the data, I make three adjustments. First, I cannot observe real applicants, only the K^F fictional experimental applicants and their data $\sum_{k\in F,k\neq i} T_{kj}$.

$$E\left[P_{ij}\middle|\boldsymbol{T_j}, \boldsymbol{u_j}\right] = \mu(1+\Gamma_0\Gamma_1)T_{ij} + \mu\Gamma_0\left[\frac{\sigma_{(T)\omega}}{\sigma_{T\omega}}\sigma_T^2 - \sigma_{(T)T}\right]\sum_{k\in F,k\neq i} T_{kj} + u_{ij} + cst. \quad (A.4.)$$

The experiment balances the contribution of real applicants to pool quality, $\sum_{k\in R,k\neq i} T_{kj}$, so this portion of $\sum_{k\neq i} T_{kj}$ can be grouped safely into the constant when estimating differences in means. Estimating $(A.4.)$ estimates the same parameters as $(A.3.)$. Hence, in the main text I suppress any indication of only using fake applications. However, the intensity of spillovers does depend separately on the number of real applicants. Provided the employer believes that T_{ij} is correlated across applicants to the same job, the denominator of Γ_0 depends on the total number of applicants, K, including real applicants. Thus, Γ_0 and the entire spillovers coefficient falls as the applicant pool rises. Intuitively, a small number of fictional applicants will have difficulty affecting the employer's assessment of the entire applicant pool when mixed with many real applicants. In the main text, I provide some evidence that applicant pools are sufficiently small to generate empirically important informational spillovers in this model.

Second, I estimate all models using the mean rather than the sum of applicant pool characteristics:

$$E\left[P_{ij}\middle|\boldsymbol{T_j}, \boldsymbol{u_j}\right] = \mu(1+\Gamma_0\Gamma_1)T_{ij} + \mu\Gamma_0\left[\frac{\sigma_{(T)\omega}}{\sigma_{T\omega}}\sigma_T^2 - \sigma_{(T)T}\right](K-1)\bar{T}_{(i)j} + u_{ij} + cst. \quad (A.5.)$$

Studies sometimes non-randomly vary the number of applications sent to a particular vacancy. Using the mean rather than a sum ensures that I do not identify spillovers off non-random variation in the number of applications. Means can also be more easily interpreted. A prior version of this paper used sums while controlling for the number of applications with very similar results.

Third, in the data we observe interview requests rather than expected productivity. Suppose that an employer faces a constant cost c of interviewing an applicant. Define Y_{ij} as an indicator of whether employer j requests an interview with applicant i. If the employer interviews any applicant demonstrating positive expected net benefits,[19] the probability of observing a callback is:

$$\Pr\left[Y_{ij}=1\middle|\boldsymbol{T_j}\right]=\Pr\left[\tilde{\alpha}+\mu(1+\Gamma_0\Gamma_1)T_{ij}+\mu\Gamma_0\left[\frac{\sigma_{(T)\omega}}{\sigma_{T\omega}}\sigma_T^2-\sigma_{(T)T}\right](K-1)\bar{T}_{(i)j}+u_{ij}\geq c\right]$$

If u_{ij} has a symmetric distribution $F(\cdot)$, then

$$\Pr\left[Y_{ij}=1\middle|\boldsymbol{T_j}\right]=F\left[\alpha+\beta T_{ij}+\delta\beta(K-1)\bar{T}_{(i)j}\right] \qquad (A.6)$$

where $\beta=\mu(1+\Gamma_0\Gamma_1)$, $\alpha=\tilde{\alpha}-c$, and $\delta=\frac{\Gamma_0}{1+\Gamma_0\Gamma_1}\left[\frac{\sigma_{(T)\omega}}{\sigma_{T\omega}}\sigma_T^2-\sigma_{(T)T}\right]$. Note that equation (A.6) corresponds to the reduced form empirical model in equations (5) and (6) if $F[\cdot]$ is linear. The coefficient on the observed characteristic measures a combination of taste-based and statistical discrimination. The coefficient on average characteristics of other applicants measures the strength of spillovers. Even in a very simple model, the tendency of employers to fully use all available information can generate positive spillovers across applicants.

Negative spillovers can also be generated if employers face capacity constraints. In the basic model, the employer faces no congestion. The interviewing cost is constant; thus the only interaction between applicants is informational. If employers instead face increasing costs of interviewing the marginal applicant, interview space becomes scarce and applicants will compete based on their characteristics. For concreteness, suppose that marginal costs rise very quickly such that the employer schedules at most one interview.[20] In addition to providing net positive benefits to the employer, an applicant must be the best applicant:

$$\beta T_{ij}+\delta\beta(K-1)\bar{T}_{(i)j}+u_{ij}\geq\beta T_{kj}+\delta\beta(K-1)\bar{T}_{(k)j}+u_{lj}\quad\forall\ k\neq i$$

$$\Leftrightarrow\beta(1-\delta)\left(T_{ij}-T_{kj}\right)+u_{ij}\geq u_{kj}\ \ \forall\ k\neq i$$

Applicants are in competition. Improving k's application will make i less likely to be the best applicant, and applicant k displaces applicant i.

Thus, the overall probability of receiving a callback will depend on whether the negative displacement effect outweighs the positive information spillover. Formally, the probability of obtaining a callback is

[19] This implicitly models the employer as choosing between hiring the applicant and not hiring at all where not hiring has a value normalized to zero. An identical but re-labelled model could allow the second option to be delaying hiring, hiring through an alternative source, or any other secondary option with a value not dependent on the data.
[20] Any number n less than the total number of applicants gives similar comparative statics.

$$\Pr\left[Y_{ij}=1\middle|\boldsymbol{T_j}\right]=F\left[\alpha+\beta T_{ij}+\delta\beta(K-1)\bar{T}_{(i)j}\right]\prod_{k\neq i}\Pr\left[\beta(1-\delta)\left(T_{ij}-T_{kj}\right)+u_{ij}\geq u_{kj}\right] \quad (A.7)$$

The effect of a second applicant on the original applicant depends on whether the positive information effect operating through the first term dominates the negative displacement effect in the second term. Finally, note that equation (A.7) corresponds to equation (1) in the main text. A simple model with correlated unobservable quality and congestion in interview schedules can rationalize either positive or negative spillovers across applicants.

Appendix 2. Equivalence of Reduced Form and Instrumental Variables Specifications

Consider the reduced form specification as in (6) where I have omitted X_{ij} for expositional simplicity.

$$Y_{ij}=\alpha+\psi Z_{ij}+\delta\psi(K-1)\bar{Z}_{(i)j}+\epsilon_{ij} \qquad (A.8)$$

Summing over all K applications to the same job:

$$\sum_{all\ k}Y_{ij}=\alpha*K+\psi\sum_{all\ k}Z_{ij}+\delta\psi(K-1)\sum_{all\ k}Z_{ij}+\sum_{all\ k}\epsilon_{ij} \qquad (A.9)$$

Subtracting (A.8) from (A.9) and re-arranging

$$\sum_{k\neq i}Y_{ij}=\alpha*(K-1)+\delta\psi(K-1)Z_{ij}+\psi[1+\delta(K-2)]\sum_{k\neq i}Z_{ij}+\sum_{k\neq i}\epsilon_{ij} \qquad (A.10)$$

Re-arranging:

$$\psi(K-1)\bar{Z}_{(i)j}=\left[\frac{1}{[1+\delta(K-2)]}\right]\left[\sum_{k\neq i}Y_{ij}-\alpha*(K-1)-\delta\psi(K-1)Z_{ij}-\sum_{k\neq i}\epsilon_{ij}\right] \quad (A.11)$$

Substituting (A.11) into (A.8) and grouping like terms:

$$Y_{ij}=\left[\alpha-\frac{\delta\alpha*(K-1)}{1+\delta(K-2)}\right]+\psi\left[1-\frac{\delta^2(K-1)}{[1+\delta(K-2)]}\right]Z_{ij}+\left[\frac{\delta(K-1)}{[1+\delta(K-2)]}\right]\bar{Y}_{ij}$$
$$+\left[\epsilon_{ij}-\left[\frac{\delta}{[1+\delta(K-2)]}\right]\sum_{k\neq i}\epsilon_{ij}\right] \qquad (A.12)$$

Three features are worth noting. First, (A.12) is identical to the instrumental variables specification defined in the main text with $\delta_{IV}=\frac{\delta}{[1+\delta(K-2)]}$. The original spillovers coefficient and the IV spillovers coefficient will have the same sign given the assumption that $|\delta|<\frac{1}{K-1}$, i.e. so long as direct effect of a characteristic on the individual holding that characteristic is greater than the

spillover effect on other applicants. Given this assumption, the spillover coefficients in the reduced form and IV specifications test the same hypothesis. Second, $\bar{Z}_{(i)j}$ is a valid instrument for $\bar{Y}_{ij}$ as long as it was exogenous in the original reduced form. Third, the coefficient on Z_{ij} will be of the appropriate sign but attenuated toward zero.

Appendix 3. Data Description for Existing Studies

Six other studies provide data which can be used to mirror the analysis of Bertrand and Mullainathan (2004). Appendix Table 1 describes the relevant details of these datasets for the present study. Booth, et. al. (2011) studies employer discrimination among Aboriginal, Anglo, Chinese, Italian, and Middle Eastern names in Australia. For comparison with other studies, I will consider treatment in this study to be an Aboriginal, Chinese, or Middle Eastern name. They partially stratify randomization of names, randomly choosing 4 out of the 5 categories for each vacancy. Eriksson and Rooth (2014) focus on discrimination against the long-term unemployed in Sweden, but their experiment includes a name treatment that they indicate will be used in other studies. Each job receives three applications with different name treatments: male native name, female native name, and foreign male name. For comparison, I will consider the foreign name as the main treatment. I exclude jobs receiving only 1 application (sent outside Stockholm and Gothenburg) as I cannot test for spillovers. Lahey (2008) does not vary ethnicity of applicant names and focuses on discrimination based on age as signaled by high school graduation date for applicants in the US. Each job receives two applications, and treatment is stratified less starkly, requiring only that the ages of the two applicants to the same job not be equal. Again, I eliminate a small number of jobs to which only one application was sent. Oreopoulos (2011) studies discrimination against high-skill job applicants in Toronto and Montreal listing immigrant names. He also studies the many sources of discrimination against immigrants by varying the location of education and experience for those with immigrant names. To keep the analysis similar to other studies, I focus on the treatment effect measuring the difference between his "type 0" and "type 1" applications which differ only due to foreign names. I include and control for his other treatments in my analysis but do not report the measured effects. I eliminate a small number of vacancies which only receive one application. I also include Phillips (2015). This study does not yet have publicly available data, but I have access to the data and thus use it. This study focuses on how low-wage employers in Washington, DC respond to the address listed on the job application but also includes a name treatment using the same assignment of "black names" as Bertrand and Mullainathan (2004). Finally, I augment these existing studies with a dataset from an experiment I

conducted on discrimination in housing. This study assigns white and black names to e-mail inquiries for apartments in Washington, DC. The primary goal of this experiment was unrelated to the present paper; however, I was able to incorporate some treatments directly related to spillovers. I randomly vary both the number of applicants to each apartment and whether name assignment is stratified. For the main analysis, I only consider apartments receiving more than one application. For greater detail on the first six studies, see the cited papers. For greater detail on the housing study, see a full explanation of the experimental design in the next appendix. Altogether, I can implement my empirical strategy using 7 different datasets spanning 15 years, 4 countries, and various types of housing and labor markets.

The empirical framework summarized in equations (6) through (8) requires that in each study I identify instrumental variables for the number of callbacks to other applications. Instruments must be assigned randomly but not stratified by job. As shown in Appendix Table 1, each study provides at least one such instrument.[21] In some studies, such instruments are easy to implement. Phillips (2015) randomly assigns work experience, age of applicant, and current unemployment duration independent of all other job and applicant characteristics. Other studies require additional control variables. For instance, Bertrand and Mullainathan (2004) assign addresses randomly conditional on the city of the job vacancy and female names randomly conditional on whether the job is a sales or administrative job. Similarly, Eriksson and Rooth (2014), Oreopoulos (2011), and the new housing study each randomly assign a large number of variables conditional on similar occupational, location, and/or experiment phases dummies. In some studies the treatment itself is the only instrument available. Both Booth, et. al. (2011) and Lahey (2008) require that applicants to the same job have different values for the treatment but otherwise select treatment randomly from a set of potential values greater than the number of applicants per job. Booth, et. al. (2011) randomly pick 4 out of 5 ethnicities to send to each job. Lahey (2008) randomly picks 2 out of 5 ages to send to each job. Thus, both studies set treatment randomly and only partially correlate treatment across applicants to the same job, which makes the treatment variable also a valid instrument. Finally, for studies in which the number of applications per job can vary, I control for the number of applications sent to ensure that I identify spillovers off variation in the instrument Z_i rather than the number of applications, which is not set randomly except in the housing study. Most importantly,

[21] Some studies could not be included in my analysis for lack of an instrument (e.g. Arceo-Gomez and Campos-Vazquez, 2014)

for each of the 7 studies listed I have at least one candidate instrument which can be used to implement the empirical strategy described above.

Careful attention to the details of each experiment's design allows me to select exogenous and non-stratified instrument candidates; however, I also wish to select instruments that are important to employers. As described above, I apply a LASSO estimator to equation (9) to select the instrument with the strongest relationship between the value of instrument and callbacks to that same application. In Appendix Table 1, I list the selected instrument in italics. The strongest instrument varies across studies. In alphabetical order of the studies, the instruments are the log median household income of the applicant's listed address, a dummy for having an Anglo name, a dummy for an extroverted personality, the age of the applicant, a dummy for listing the name "Carrie Martin," the age of the applicant, and a dummy for including the statement "Just so you know, I am a smoker and my credit rating is below average" in the e-mail inquiry for housing. These variables have been identified empirically as important to employers. In the above text, I will test whether changing the values of these instruments for one applicant spills over to other applicants.

Appendix 4. Description of Housing Experiment

In addition to existing datasets, I occasionally make use of a correspondence experiment that I conducted in the housing market. The housing market provides a useful contrast to the labor market as landlords' ability to adjust the number or timing of apartments for rent may be less than employers' ability to adjust the number or timing of job vacancies. At the same time, housing markets involve a similar search process with one side of the market posting vacancies and the other side applying to these vacancies. Thus, finding similar results in housing would suggest that positive spillovers are a general feature of search rather than just a feature of labor markets. Finally, housing markets provide a convenient context in which to complete a correspondence experiment as in Hanson and Hawley (2011) and Ewens, et. al. (2014). However, these previous studies have generally sent only one application per vacancy or have not made data publicly available. As such, I make use of a new correspondence study of the housing market in Washington, DC. This experiment was completed primarily for reasons unrelated to the present study but allows for some additional experimentation.

During May and June 2015, a research assistant applied to apartment vacancy listings from an online classified ad site. Apartments were randomly chosen out of those listed within the previous 24 hours in the District of Columbia to which the experiment had not applied previously,

which were not obviously a scam, and which were monthly rentals. For the purposes of a separate study, the sample focuses on housing for low-income individuals. Apartments were restricted to those with rent below $1,500 per month and those that provided the location of the apartment. During the first half of the experiment, apartments were randomly selected to receive 1 or 2 applications. During the second half, this increased to 2 or 4 applications. Altogether, the experimental sample includes 2,681 applications to 1,342 apartments.

An inquiry to a particular apartment included a randomly generated message largely using the same content as Hanson and Hawley (2011) and Ewens, et. al. (2014). Appendix Figure 1 displays an example of such a message as well as the general template. Each e-mail includes a subject followed by a message consisting of a greeting, introductory statement including the applicant's name, a request regarding the availability of the apartment, and a valediction finishing with the applicant's name. Names are chosen at random from the same list as Bertrand and Mullainathan (2004). For comparison later, I randomly choose some apartments to have equal number of names of each race (i.e. stratified randomization) and other apartments to have the race of each name drawn independently. As in Ewens, et. al. (2014) I also randomly and independently assign some applicants to include positive quality signals (professional employment, good references, and/or good credit) or negative signals (smoker and/or bad credit) and others to have no signal statement. Given the focus on lower-rent housing, I also randomly and independently assign applicants to include a request about whether the apartment accepts Section 8 housing vouchers. To avoid sending the same exact wording of a particular component to the same apartment, I compose 4 possible versions of each element. During the first phase, each apartment received up to 2 messages randomly chosen without replacement from 2 possible versions of each element. During the second phase, messages were randomly chosen from 4 possible versions without replacement. Once composed, a research assistant sends the messages from e-mail accounts matching the applicant's name.

As in Ewens et. al. (2014), I focus on only positive responses, and I categorize a response as positive if the landlord invites the applicant to setup a showing of the apartment, explicitly provides a means for further contact (e.g. asks to call a particular phone number), or responds that the apartment is available while providing or requesting more information. I do not include negative responses, primarily those stating that the unit is no longer available or that some stated trait of the applicant is incompatible with the apartment ("no Section 8" or "no smokers"). Following Ewens et. al. (2014), I also do not include "disinterested" landlords who provide/request more information

without answering whether the apartment is available and landlords who simply state that the apartment is available and nothing else.

I apply the same instrumental variables framework as above (equations (7) and (8)) to this housing experiment to test for positive spillovers and bias. As before, I will consider discrimination based on the apparent race of the applicant's name the main treatment of interest. I have many potential instruments available which are randomly assigned, affect the landlord's response to a given message, and are not fully stratified. To use all of this potential variation in other applicants' quality, I construct dummies for whether any other applicant uses each possible greeting, introduction, quality statement, section 8 statement, and valediction as well as the total number with black names and total number with female names. These variables are all randomly assigned. Additionally, whether other applicants to the same apartment list a black name, list a female name, request Section 8, or provide a positive or negative quality signal are not fully stratified. The specific messages are partially stratified (two section 8 requests with the same text are not sent to the same landlord). However, the specific content is sometimes chosen from a pool of potential sentences that is larger than the number of applications sent, providing variation in the applicant pool for a given housing vacancy. As with existing studies these variables are randomly assigned but only conditional on a small number of covariates. I control for a dummy for the second phase of the experiment (during which the pool of potential messages to be sent was expanded) and a set of dummies for the number of applications sent to a given vacancy.

Appendix 5. Simulating Variable Selection on Known Weak Instruments

In theory, the strength of the quality instrument could remain suspect due to the variable selection process; however, the available evidence allays this concern. As described in the main text, a study sometimes provides multiple exogenous instruments from which I select the strongest instrument. With a sufficiently large number of candidate instruments, the strongest instrument will provide large first-stage F-statistics in a given sample even if each candidate instrument has no first stage relationship in the population. However, some simulations alleviate this concern that I mechanically generate a strong instrument. I simulate whether random noise instruments can generate a first-stage F-statistic approaching 80 in the context of the present study. For each simulation, I create a new 'scrambled' dataset. I keep the outcome data Y_{ij} and its group structure vis-à-vis the job unchanged. I then randomly select (with replacement) the independent variables from some person m in the same study sample, $[X_{mj}, Z_{mj}]$, to pair with the outcome data of person i. Using this scrambled data, I select the strongest instrument and replicate the first stage from

Table 2. The first stage now measures the relationship between person i's outcome and randomly chosen person m's covariates. This "scrambled" first stage has a no true relationship by construction. Person i and person m have no systematic connection. Hence, a strong first stage could only result from selecting the instrument with the strongest chance relationship out of multiple noise variables. Appendix Figure 2 displays the distribution of first stage F-statistics generated by simulating this scrambled first stage 1,000 times. The variable selection process sometimes generates F-statistics above 10 from truly weak instruments; however, it never generates a first stage F-statistic above 40, let alone the value of 77.8 from the actual data. While variable selection could mask weak instruments, the particular process I use does not present this problem.

Appendix 6. Simulations and Estimation of MSE

To calculate mean squared error I need to quantify bias and variance of estimated treatment effects. I estimate bias using omitted variable bias calculations from the main text. I summarize these in the first column of Appendix Table 2, which matches bias in Table 6. The second column of Appendix Table 2 then squares these differences as in the formula for MSE and multiplies the result by 1,000 for readability. For example, a simple OLS model in Bertrand and Mullainathan (2004) estimates treatment effects that are 1.9 percentage points too small. The first row of Appendix Table 2 takes the simple average across the six[22] studies with a racial name treatment. On average, these studies underestimate treatment effects by 2.0 percentage points. The square of bias (x1000) averaged across studies then contributes 1.01 to MSE (x1000). On the other hand, non-stratified designs eliminate this bias; hence, I list a bias of zero for non-stratified designs.

To measure the efficiency gains from stratified designs, I conduct a set of simple Monte Carlo simulations that generate a bootstrap measure of variance. To simulate a stratified design for a particular study, I generate and randomly assign with stratification a fake treatment dummy, $\tilde{T}_{ij}$, to each observation in the original data. For vacancies with 4 applicants in the original data, I randomly assign exactly 2 to be "treated." For 2 or 3 applicants, I assign exactly 1 to be "treated." To simulate a non-stratified design, I simple assign treatment randomly and independently across all applicants in the sample and choose the probability of treatment so that it equals that of the stratified

[22] Again excluding Lahey (2008) because the age treatment generates effects with magnitudes that cannot be directly compared to race treatments.

simulation.[23] In either case, I then run a simple OLS estimate of treatment effects using actual outcome data, actual control variables, and the new fictional treatment:

$$Y_{ij} = \alpha + \beta \tilde{T}_{ij} + \psi Z_i + \zeta X_i + \epsilon_{ij} \quad (A.13)$$

I repeat this procedure 1,000 times, each time drawing a new sample with replacement, clustering by vacancy, and drawing the same number of vacancies as in the original data. I measure $\hat{\beta}$ across these repetitions and calculate its sample variance.[24]

The third column of Appendix Table 2 shows the results of these simulations. As expected, stratification improves precision by guaranteeing baseline balance on vacancy characteristics. For example, in the Bertrand and Mullainathan (2004) study, the variance of the estimates (x1000) contributes 0.06 to MSE in a non-stratified design but only 0.04 in a stratified design. Across the six studies with racial name treatments, stratifying treatment by vacancy decreases the variance (x1000) of the estimated treatment effects from 0.18 to 0.11, a decrease of 40%. However, the bias generated by stratification in the presence of spillovers outweighs efficiency gains. The fourth column of Appendix Table 2 shows this result. A stratified design multiplies MSE (x1000) 6 times in the Bertrand and Mullainathan (2004) study from 0.06 to 0.40. While stratifying treatment lowers the variance of measured treatment effects from 0.06 to 0.04, the increase in squared bias from 0.00 to 0.36 easily outweighs the efficiency gains. Likewise, taking the mean across six studies, MSE (x1000) increases from 0.18 to 1.11.

Finally, I explore whether stratified designs can outperform non-stratified designs for small-sample studies. As the sample size shrinks, stratification generates greater efficiency gains, but bias does not vary with the sample size. I implement the same Monte Carlo simulations drawing samples of 500, 300, 100, and 50 vacancies from the original datasets rather than sample sizes equal to those from the original studies. All of these candidate sample sizes are smaller than the average (1,935) and minimum (565) actually used in the original studies. Figure 2 displays the results graphically. For the Bertrand and Mullainathan (2004) data, the non-stratified design continues to out-perform the stratified design for 500, 300 and 100 jobs, but efficiency gains from stratifying dominate for samples less than 100 jobs. Similar results are true averaging across studies. The right hand pane of Appendix Table 2 quantifies these results.

---

[23] Suppose the original study sends 4 applicants to half of the jobs and 3 applicants to the other half. The stratified simulation will assign 2 of 4 and 1of 3 applicants to the fake treatment. Hence, the overall probability of treatment for the non-stratified simulation should be $\frac{1}{2} * \frac{1}{2} + \frac{1}{2} * \frac{1}{3} = \frac{5}{12}$ to ensure comparability.

[24] Because the "treatment" does not involve any actual interaction with the world, the average of $\hat{\beta}$ across the simulations is zero.

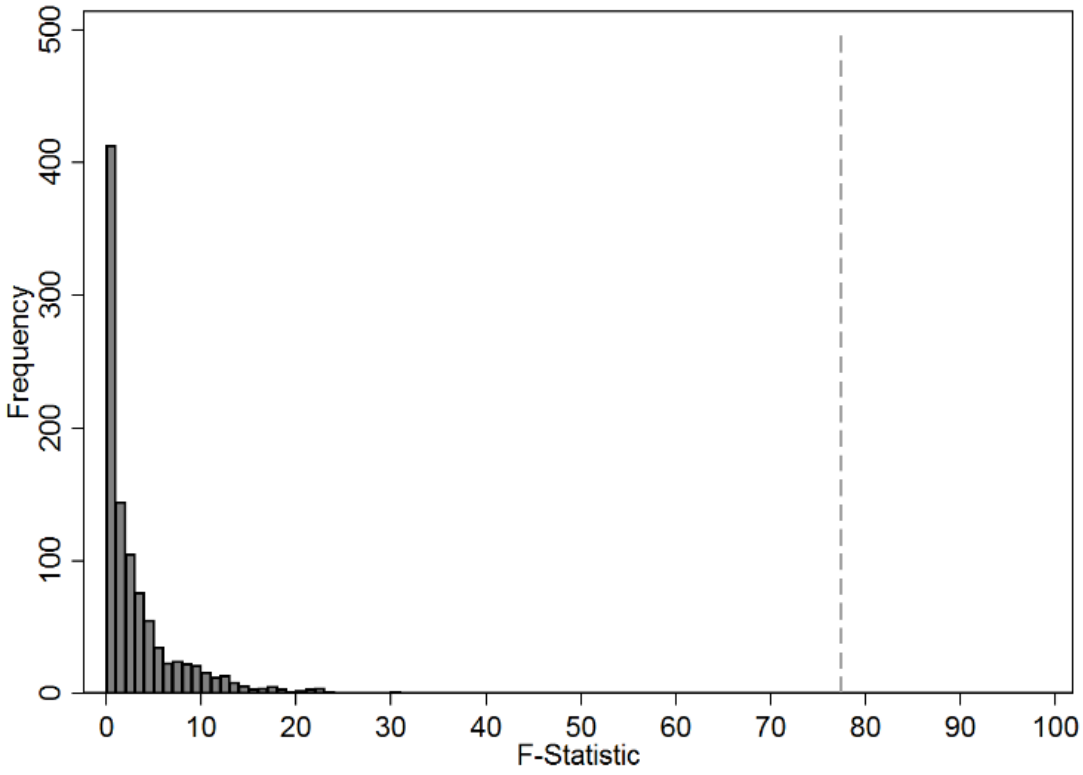
Appendix Figure 1: Message Example and Template for Housing Study

Subject: Interested in Your Craigslist Ad
Dear Sir:
My name is Latoya Williams, and I saw the place on the internet for RENT AMOUNT/month. If you need them, I have good references and I could also send a recent credit report. I would also like to know if you accept Section 8 vouchers. Is the place still available?
Sincerely,
Latoya Williams

……
……

Subject: «Subject»
«Greeting»
«Introductorty Statement Including Name» «Quality Statement» «Section 8 Statement»
«Availability Question»
«Valediction»
«Name»

Appendix Figure 2: First Stage F-Statistics from Simulated Weak Instruments Relative to Actual Value



Notes: The histogram shows first stage F-statistics from 1,000 simulations. Each simulation regresses a first stage equation as in equation (7). Instead of using the actual data, the strongest instrument is chosen from noise variables with no true first stage. I generate noise variables by matching the outcome data of one applicant to the independent variables of a different, randomly-chosen applicant. The dashed line provides the actual first-stage F-statistic for comparison.

Appendix Table 1. Description of Existing Labor Market Experiments

Study	Treatment Variables	Treatment Stratified by Job?	Applications per Vacancy	Sample Size	Sample Restrictions	Candidate Instruments/*Selected Instrument*	Control Variables
Bertrand and Mullainathan (2004)	Black name	Yes	2 or 4	4,546/4,870	Only jobs where all applications have neighborhood income values	*Log median household income of listed address*; female name	City dummy, sales job dummy, number of applications per job dummy
Booth, et. al. (2011)	Foreign name	Somewhat (3 or 4 of 5)	3 or 4	4,212/4,212	None	Name ethnicity dummies (*Anglo dummy*); name sex	Name ethnicity
Eriksson and Rooth (2014)	Foreign name; female name	Yes	3	6,873/8,466	Only jobs with more than one application sent	Work history variables, personality characteristics dummies (*extroversion dummy*), leisure activity dummies, US high school dummy, more education than required dummy, summer job dummy	Occupational dummies, regional dummies
Lahey (2008)	Age	Somewhat (ages cannot be equal)	2	7,932/8,002	Only jobs with more than one application sent	*Age*	Age
Oreopoulos (2011)	Foreign name	Yes	2, 3, or 4	12,906/12,910	Only jobs with more than one application sent	Bachelor's degree quality, language skills dummy, master's degree dummy, high quality experience dummy, Canadian reference dummy, legal dummy, education accreditation dummy,	Experiment phase dummies, city dummy, treatment type dummies, foreign education dummy, number

						extracurricular activities dummy, set of dummies for each specific name (*name = "Carrie Martin"*)	of applications per job dummy
Phillips	Black name	Yes	4	2,260/2,260	None	Years of work experience, *age*, current unemployment duration	None
New Housing Study	Black name	Sometimes	1, 2, or 4	2,220/2,681	Only jobs with more than one application sent	Black name, female name, subject dummies, and dummies for all possible greeting, introduction, apartment request, valediction, quality signal (*smoker AND low credit message*), and section 8 statements	Experimental phase dummy; number of applications dummy

Appendix Table 2. Simulating the Tradeoff between Bias and Efficiency in Stratified Designs

Number of Clusters:		All Clusters				500	300	100	50
	Statistic Displayed:	Bias	Bias Squared x 1000	Variance x 1000	MSE x 1000	MSE x 1000	MSE x 1000	MSE x 1000	MSE x 1000
Average	Not Stratified	0.000	0.00	0.18	0.18	0.39	0.69	2.01	3.93
	Stratified	0.020	1.01	0.11	1.11	1.26	1.42	2.30	3.30
Bertrand and Mullainathan	Not Stratified	0.000	0.00	0.06	0.06	0.15	0.26	0.81	1.45
	Stratified	0.019	0.36	0.04	0.40	0.45	0.51	0.82	1.24
Booth et. al.	Not Stratified	0.000	0.00	0.20	0.20	0.48	0.85	2.37	4.72
	Stratified	0.001	0.00	0.12	0.12	0.32	0.50	1.68	2.86
Eriksson and Rooth	Not Stratified	0.000	0.00	0.11	0.11	0.53	0.97	2.71	5.07
	Stratified	0.072	5.18	0.07	5.25	5.50	5.69	6.92	8.28
Lahey	Not Stratified	0.000	0.00	0.02	0.02	0.19	0.32	0.93	1.68
	Stratified	0.000	0.00	0.01	0.01	0.05	0.08	0.26	0.59
Oreopoulos	Not Stratified	0.000	0.00	0.03	0.03	0.17	0.32	0.92	1.90
	Stratified	0.002	0.00	0.02	0.02	0.12	0.20	0.56	1.10
Phillips	Not Stratified	0.000	0.00	0.28	0.28	0.29	0.49	1.48	3.16
	Stratified	0.022	0.48	0.17	0.65	0.66	0.79	1.38	2.22
Housing	Not Stratified	0.000	0.00	0.41	0.41	0.74	1.27	3.78	7.26
	Stratified	0.003	0.01	0.27	0.28	0.52	0.82	2.44	4.12

Each pair of rows corresponds to a different dataset. The "stratified" row shows bias from Table 6, and the "non-stratified" row has bias of zero by definition. The variance of the estimates comes from 1,000 Monte Carlo simulations randomly generating a fake treatment dummy (either stratified or not) and regressing actual callback outcomes on this treatment dummy. In the left side of the table, I use a full sample drawn with replacement from the original data clustering by vacancy. The right side of the table draws samples of smaller size where the number of clusters corresponds to the number of jobs drawn in the sample.